

Ibrahim, Musadiq (2013) Probing function of unknown proteins by using pharmacophore searching and biophysical techniques. PhD thesis.

<http://theses.gla.ac.uk/3924/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

# Probing Function of Unknown Proteins by using Pharmacophore searching and Biophysical techniques



Musadiq Ibrahim

Submitted in fulfilment of the requirements for the degree of Doctor  
of Philosophy

School of Chemistry  
College of Science and Engineering  
University of Glasgow

January 2013



## Abstract

The number of protein structures deposited in the Protein data bank is increasing almost exponentially and among these structures many of the proteins are novel with unknown function. Like Docking, Pharmacophore searching is an *In-silico* technique which is widely used for drug discovery. In pharmacophore searching the main focus is on the hydrogen bond interactions between the ligand and the target protein. The pharmacophore models are generated either by using the already known actives as templates or by utilizing the significant chemical features of the active site.

In this thesis the pharmacophore searching has been used to find potential ligands/substrates for unknown proteins and then ligand binding is confirmed by using different biophysical techniques. In the initial phases the pharmacophore models were generated by using Cerius2 and Weblab Viewer pro programs. While in later stages more sophisticated searches were carried out by using DSV (Discovery studio visualizer, Accelrys®). Procedures were optimized for model building by using DSV, which enabled the pharmacophore searching via both the Vector and the Query atom methods. To validate the technique, it was first used on known enzymes with established function e.g. xylose reductase and shikimate kinase. The optimized pharmacophore model when search through the database successfully identified the true substrates for these enzymes among other ligands thereby demonstrating the attainment.

In addition, protein structures from the protein data bank (PDB) with unknown ligands (UNK) were identified and manually screened to find examples that could be used to test the applicability of pharmacophore searching methods. The diversity of structures showed that the definition of an unknown ligand is completely inconsistent with many examples where any non spherical density was labelled as unknown ligand and in most cases a single atom is labelled as an unknown ligand, which most likely can be an ion or a water molecule. It appeared that some compounds like glycerol, phosphate and citrate which co-crystallized with the protein due to their presence in the crystallization conditions were also mistakenly assigned as UNK. The pharmacophore method worked successfully in finding suitable ligand (s) for the protein.

The technique has been used to find potential ligands for proteins with unknown function on three test cases e.g. TdcF, HutD and PARI. Of the potential pharmacophore hits obtained through database search, a number of compounds were either purchased or synthesised to be tested for binding affinity. Different biophysical techniques like DSC, ITC, CD and NMR were used for this purpose. Among these techniques NMR proved to be the most sensitive technique to differentiate binders from non-binders and to further detect weak and strong bonding in terms of  $K_d$  values. For TdcF among other binders the best binder was 2-ketobutyrate with a  $K_d$  value of 200 $\mu$ M. In case of HutD, formyl glutamate ( $K_d$  = 92 $\mu$ M) and formimino glutamate ( $K_d$  = 500 $\mu$ M) came out to be the best binders and could be the true ligands of the protein at physiological concentration. For PARI L-glutamate appeared to be a potential ligand for the protein as confirmed through the NMR experiments. Pharmacophore modelling has been successful in identifying potential interactions provided by the protein active site which in turns specifies the required features to be present in a ligand and later on the successful binding studies further confirm its applicability.

While developing pharmacophore methodology and performing binding experiments, a number of other projects were also carried out. In a side project enzyme inhibition assays were carried out on an enzyme called NFGase. The  $K_i$  values were calculated for different inhibitors along with the type of inhibition. Among different substrate analogues, one compound named 1,2,4 butane tricarboxylic acid showed inhibition in the nano molar range ( $K_i$  = 190nM).

The enzyme characterization of different DHQases was carried out and the X-ray structure of *Campylobacter jejuni* DHQase along with 3-Dehydro shikimate was determined at a resolution of 2.4Å. The enzyme assays were carried out for the first time for *Campylobacter jejuni* & *Candida albicans* DHQase, which gave the lowest  $K_m$  (17 $\mu$ M) value for the former. The comparison of the 3 available X-ray structures (in complex with 3-Dehydro shikimate) indicated that the presence (*Campylobacter jejuni* DHQase, *Mycobacterium tuberculosis* DHQase) and absence (*Streptomyces coelicolor* DHQase) of Gly78 residue in the active site suggests reasons for the observation of low and high  $K_m$  values among different DHQases.

## Table of Contents

Author's Declaration .....	I
1. INTRODUCTION .....	1
1.1 Protein from Structure to Function .....	1
1.1.1 Protein Sequence comparisons .....	2
1.1.2 Protein Fold similarity .....	4
1.1.3 Protein functional Assignment .....	8
1.1.4 Protein-ligand binding Interactions .....	9
1.1.5 Solvent molecules and protein binding site .....	12
1.2 <i>In silico</i> Techniques for functional annotation of Novel proteins .....	13
1.2.1 Docking .....	13
1.2.2 Pharmacophore .....	18
1.2.3 Pharmacophore Searching .....	20
1.2.4 Pharmacophore generating programs .....	23
1.2.5 Visualization of Hits .....	24
1.2.6 Hits Screening .....	25
1.2.7 Future of Pharmacophore Searching .....	25
1.3 Biophysical techniques for binding studies .....	26
1.3.1 X-ray Crystallography .....	26
1.3.2 Nuclear magnetic resonance (NMR) .....	32
1.3.3 High throughput screening (HTS) .....	39
1.3.4 Isothermal titration calorimetry (ITC) .....	41
1.3.5 Differential scanning calorimetry (DSC) .....	42
1.3.6 Circular Dichroism (CD) spectroscopy .....	43
1.4 Aims and objectives of the project .....	44
2. MATERIALS AND METHODS .....	45
2.1 General reagents .....	45
2.2 Bacterial strains .....	45
2.3 Plasmids .....	46
2.4 Plasmid purification (mini prep) .....	47
2.5 DNA quantification .....	48
2.6 pH measurements .....	48
2.7 Antibiotics: .....	48
2.8 Culture media for bacterial growth .....	49
2.9 Preparation of ultra-competent E.coli (BL21 DE3) cells: .....	50
2.10 Transformation protocol .....	51
2.11 Storage of bacterial Strains .....	51
2.12 Protein over-expression .....	52
2.12.1 Overnights and test of expression .....	52
2.12.2 Large scale growth of bacterial cultures .....	52
2.12.3 Cells harvesting .....	52
2.12.4 Cell lysis by ultrasonication .....	53
2.13 Protein purification .....	53
2.13.1 Ni-NTA affinity chromatography or IMAC .....	53
2.13.2 Gel filtration chromatography .....	54
2.14 Dehydroquinases purification protocol .....	54
2.14.1 Ion exchange chromatography .....	55
2.14.2 Hydrophobic interaction chromatography (HIC) .....	55
2.14.3 Gel filtration chromatography on superdex 200 column .....	55
2.15 Protein characterization .....	56
2.15.1 SDS-PAGE .....	56
2.15.2 Sample preparation for SDS Gel Electrophoresis .....	56
2.15.3 Staining and destaining procedure .....	56

2.15.4	Measurement of protein concentration via UV absorbance .....	57
2.15.5	Dialysis .....	57
2.15.6	PD-10 desalting column (Buffer exchange) .....	57
2.15.7	Lyophilization (Freeze drying).....	57
2.16	Protein crystallization .....	58
2.16.1	Streak seeding .....	59
2.17	Circular dichroism (CD) spectroscopy.....	59
2.18	Isothermal titration calorimetry (ITC) and differential Scanning calorimetry (DSC) .....	59
2.19	Enzyme Assays .....	59
2.19.1	Standard 3-dehydroquinase assay.....	60
2.19.2	Standard NFGase assay .....	60
3.	Pharmacophore methodology .....	61
3.1	Overview of pharmacophore methodology.....	61
3.2	Catalyst® .....	63
3.3	Accelrys Discovery studio visualizer® (DSV) .....	64
3.4	Generation of Databases .....	64
3.4.1	Generation of Aldehyde and ketone compounds database.....	64
3.4.2	Generation of dipeptide and tripeptide database.....	65
3.5	Pharmacophore generating methods .....	65
3.5.1	Cerius2, Notepad and WebLab viewer pro method.....	66
3.5.2	DSV, Vector method .....	69
3.5.3	DSV, Query atom method .....	82
3.6	Advantages of query atom method .....	88
4.	Xylose Reductase.....	91
4.1	Pharmacophore searching for xylose reductase.....	93
4.1.1	Pharma xylo 1.....	94
4.1.2	Pharma xylo 2.....	95
4.1.3	Pharma xylo 3.....	97
4.1.4	Pharma xylo 4.....	99
4.1.5	Manual selection of hits .....	100
4.2	Conclusions based on search models and hits .....	102
5.	Shikimate kinase <i>form Mycobacterium tuberculosis</i> .....	108
5.1	Pharmacophore searching for SK .....	109
5.2	Aims and objectives .....	109
5.3	Generation of Pharmacophore.....	110
5.4	Pharmacophore generation via holo-enzyme .....	111
5.4.1	Pharma Holo 1 .....	111
5.4.2	Pharma Holo 2 .....	112
5.4.3	Pharma Holo 3 .....	113
5.4.4	Pharma Holo 4-5.....	114
5.4.5	Pharma Holo 6 .....	115
5.4.6	Pharma Holo 7 .....	116
5.4.7	Pharma Holo 8 .....	116
5.4.8	Pharma Holo 9-10 .....	117
5.4.9	Pharma Holo 11-14.....	118
5.4.10	Pharma Holo 15-16.....	118
5.5	Apo-enzyme for pharmacophore generation.....	119
5.5.1	Pharma Apo 1 .....	119
5.5.2	Pharma Apo 2 .....	120
5.5.3	Pharma Apo 3 .....	120
5.6	Conclusions .....	122
6.	The <i>Escherica coli</i> protein <i>TdcF</i> , Pdb code: 2UYN .....	125

6.1	Introduction .....	125
6.2	Generation and Optimization of pharmacophore for <i>TdcF</i> .....	127
6.2.1	Pharma TdcF 1 .....	127
6.2.2	Pharma TdcF 2 .....	129
6.2.3	Pharma TdcF 3 .....	130
6.2.4	Pharma TdcF 4 .....	131
6.2.5	Pharma TdcF 5 .....	132
6.2.6	Pharma TdcF 6 .....	132
6.2.7	Pharma TdcF 7 .....	132
6.2.8	Pharma TdcF 8 .....	133
6.2.9	Pharma TdcF 9 .....	133
6.3	Expression and Purification of <i>TdcF</i> : .....	136
6.4	Ligand binding experiments against known ligands.....	138
6.4.1	DSC Experiments .....	138
6.5	Circular Dichroism (CD) Experiments .....	141
6.6	NMR experiments .....	144
6.6.1	NMR titrations .....	144
6.6.2	HSQC Spectra and $K_d$ determination.....	145
6.6.3	$K_d$ values of different ligands for TdcF.....	151
6.6.4	Pharma TdcF 10 (Generation of Query ligand).....	152
6.6.5	Pharma TdcF 11 .....	153
6.6.6	Pharma TdcF 12 .....	154
6.6.7	Pharma TdcF 13 .....	154
6.7	Co-Crystallizations for <i>TdcF</i> .....	155
6.8	Conclusions and Future work .....	156
7.	HutD, a protein of unknown function from <i>Pseudomonas aeruginosa</i> PAO1.....	157
7.1	Histidine utilization (Hut) pathway .....	159
7.2	Putative role of <i>HutD</i> .....	160
7.3	Aims and objectives .....	161
7.4	Over-expression and purification of HutD.....	161
7.4.1	Using Auto induction media.....	162
7.4.2	Using M9 media .....	162
7.4.3	Changing temperature and IPTG concentration of the cultures...	163
7.4.4	The effect of temperature on HutD growth .....	164
7.4.5	Harvesting and Purification of HutD.....	165
7.4.6	Size exclusion chromatography (Gel Filtration) .....	165
7.5	Pharmacophore searching for HutD .....	169
7.5.1	Pharma HutD 1 .....	170
7.5.2	Pharma HutD 2 .....	171
7.5.3	Pharma HutD 3 .....	173
7.6	Ligand selection and synthesis.....	174
7.6.1	Synthesis of N-Forminino L-Glutamate (FIGLU) .....	174
7.7	CD results for HutD .....	175
7.8	Fluorescence spectroscopy on HutD .....	178
7.9	NMR results for HutD .....	179
7.9.1	HutD with FIGLU.....	180
7.9.2	HutD with NFIGLU.....	182
7.9.3	HutD with Urocanate .....	183
7.9.4	HutD with L-Glutamate.....	184
7.9.5	HutD with L-Histidine.....	185
7.9.6	Determination of $K_d$ values .....	185
7.10	Conclusions and future work.....	186

8.	<i>Ralstonia eutropha</i> N-formylglutamate amido-hydrolase like protein (PARI)	188
8.1	Over-expression and purification of PARI .....	189
8.1.1	Lowering temperature of cultures in auto induction media. ....	190
8.1.2	Using M9 media (supplemented with zinc sulphate) .....	192
8.1.3	Purification of PARI .....	193
8.2	Protein aggregation problems .....	196
8.3	NFGase activity of PARI.....	199
8.3.1	Spasm search.....	201
8.4	Pharmacophore searching for PARI .....	204
8.4.1	Generation of pharmacophore for PARI.....	204
8.5	Ligand selection and synthesis.....	210
8.5.1	Synthesis of Carbamoyl Glutamate: .....	211
8.6	CD results for PARI .....	212
8.7	NMR experiments on PARI .....	215
8.7.1	N-formyl L-glutamate with PARI .....	217
8.7.2	Carbamoyl histidine with PARI .....	218
8.7.3	L-glutamate with PARI.....	219
8.7.4	Carbamoyl glutamate with PARI .....	220
8.7.5	Carbamoyl lysine with PARI.....	221
8.7.6	Carbamoyl cysteine with PARI.....	222
8.8	Effect of ligand addition on aggregation of PARI .....	223
8.9	Conclusions and Future work .....	225
9.	Biochemical characterization of <i>Pseudomonas auroginosa</i> Amidohydrolase (PAA) .....	226
9.1	Aims and objectives .....	227
9.2	Expression and purification of PAA.....	228
9.3	I T C studies on PAA with potential inhibitors.....	229
9.4	Inhibition studies on PAA .....	232
9.4.1	Calculation of $K_i$ for BTCA and CG .....	235
9.4.2	PAA Inhibition studies using Nickel, EDTA, BTCA and Zn solution.....	237
9.4.3	Dialysis of PAA with BTCA and Inhibition Assays.....	238
9.5	Crystallization trials for PAA with BTCA.....	239
9.6	Conclusions .....	240
9.7	Future Aims.....	240
10.	Probing protein X-ray structures with unknown ligands .....	241
10.1	Introduction .....	241
10.2	Aims and Objectives.....	242
10.3	Screening of Protein X-ray structures .....	243
10.3.1	1VKM .....	247
10.3.2	3P6K.....	248
10.3.3	2AAM .....	250
10.3.4	2F6R.....	252
10.3.5	1TVF.....	254
10.3.6	2PY6.....	256
10.3.7	3EZU.....	257
10.3.8	1VK8.....	257
10.3.9	2FTR.....	261
10.3.10	3NNR.....	262
10.4	Conclusions .....	267
11.	Type II DHQases.....	271
11.1	Aims and objectives .....	272

11.2	Expression and Purification of DHQases .....	272
11.3	Kinetic studies on DHQases.....	276
11.3.1	Enzyme Assays .....	276
11.4	Crystal structure of <i>C.jejuni</i> DHQase.....	277
11.4.1	Crystallographic Data collection and processing .....	278
11.4.2	Structure analysis .....	283
11.5	Future work.....	286
12.	<i>HISdhL (Haemophilus Influenzae SdhL)</i> .....	287
12.1	Expression and purification .....	289
12.2	Crystallization of <i>HISdhL</i> .....	295
12.3	Conclusions and future work.....	297
13.	References .....	299
	Appendices (Appendix 1) .....	311

## Acknowledgments

First of all I would express my huge gratitude to my Supervisor Dr. Adrian Lapthorn who was always available to help, discuss and guide. Without his invaluable ideas, input and feedback this project would have not been possible. I am highly indebted to Dr. Brian O Smith for giving me the opportunity to learn and perform NMR experiments and also for his generosity in the laboratory matters. I am thankful to Dr. Sharon Kelly for carrying out all the CD experiments and Margaret Nutley for all the DSC and ITC experiments. It would be unfair if I do not mention Steven Vance who was always there in the laboratory for guiding and helping despite his own busy PhD schedule. I am thankful to Dr. Nichola Picken whose facilitative role was very encouraging in the initial period of my PhD.

I am very grateful to Ross and Kate who were always there to help me through thick and thin. Either they were project students or regular members of the laboratory; I found every body fully supportive during tough times which was a huge inspiration all along. It was a privilege to interact with all of them and their unconditional help made it possible to bring this ambitious project to fruition. Big thanks to Stuart Mackay for fixing my Laptop which was often having problems.

Out side the laboratory which was not often long enough I had great time with Fida, Javed, Sajjad, Bilal, Shahid, Hainan, Hassan, Naqeeb, David, Matthias and Kashif who were a constant source of support. I would like to take the opportunity to thank all of them for their moral support and thought full company.

Finally I am thankful to Kohat University of Science and Technology (KUST) for the financial support.

Last but not the least I would like to thank my family whose prayers and support was always there, especially my Father whose guidance was invaluable at all occasions.



## **Author's Declaration**

It is declared that this thesis has been written in accordance with the University of Glasgow regulations and has not been presented for a degree at any other university. All the stated work is original unless indicated otherwise by reference in the text. The work contained within is the author's own except the work done in collaboration as indicated.

Musadiq Ibrahim©

January 2013

# 1. INTRODUCTION

## 1.1 Protein from Structure to Function

Significant research effort continues to be applied to identify the function of proteins within a given cell or organism. One of these approaches is to relate protein structure to its function by using single crystal X-ray diffraction or by nuclear magnetic resonance (NMR). But this is not simple as stated by Brian Shoichet “Even if you know what a protein looks like, this does not necessarily mean you know what it does”. To take the next step and ask “ whether we can broadly predict function of an enzyme if we know the structure” {1}. For an enzyme, substrate prediction is challenging and needs great care as said by Shoichet “But in case all you have to do is get a molecule that the enzyme recognizes. It doesn’t have to do the next step, which is to turn the molecule over, to do the catalytic reaction on it and capturing that is no joke” {1}. Various methodologies have been developed primarily for the pharmaceutical industry for the identification/design of inhibitors using the target protein structure as a template. In particular, pharmacophore searching allows hypothesis driven searching of the protein structure, but has not been applied to the protein structure/function problem due to the expense of this commercial software. Instead most approaches have used freely available academic docking software and among them one of the success case includes the prediction of activity for a protein(Tm0936) from *Thermotoga maritima* with an unknown function on the basis of docking as an enzyme {2}.

Pharmacophore searching as implemented in the available software is very ligand centric, in some cases relying on the existence of a ligand bound structure as a starting point. This reflects its roots in the drug discovery field. The number of genomes sequenced is increasing from all types of organism and with each deposited genome; there are a significant number of sequences with no similarity to any known protein. By default the function of these protein are unknown, and the percentile of such type of sequences deposited is nearly 50% {3} . The prediction of function of an enzyme on the basis of sequence similarity is the main criterion of genome annotation, and recently the correlation between the sequence and structural and functional aspects of proteins have

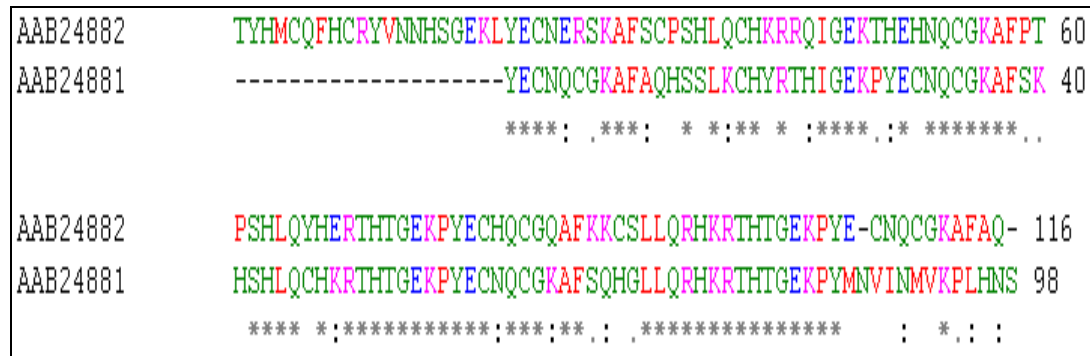
been addressed {4, 5}. However sequence similarity does not always correlate with conserved function. This suggests that there is need of different approaches for exploring the uncharacterized protein in an integrated manner and computational structural functional approaches can prove equal to the task regarding functional annotation {3}. The availability of a high resolution structure of a protein opens different avenues for analysis of an enzyme mechanism and understanding the functional role of the protein.

As the number of x-ray structures of unknown proteins deposited in the PDB is increasing hugely the focus is towards the use of various computational methods for functional annotation of these proteins {6}. Generally the active site of an enzyme is present in certain cavities or pocket like structures of the enzyme {7} and work on these type of active site has been reviewed {8} but it has been noticed that within the super families of protein there are functional differences which can be minor or quite significant in nature {6}. The interaction between the protein and ligand includes steric fit or complementarity between ligand and protein. Protein-ligand interactions can induce significant conformational changes and these physiochemical interactions involve significant solvent and solvation effects {9}. Approaches for predicting the function of a protein based on their active site includes sequence similarity, fold similarity, docking and pharmacophore searching.

### ***1.1.1 Protein Sequence comparisons***

In April 2012 there were 151,824,421 sequence records in the online database Gene Bank consisting of billions of individual amino acids/bases. With genome sequencing projects around the world this number is increasing at an exponential rate. For any specific protein its amino acid sequence can be compared against databases such as Gene Bank for similar sequences.

Sequence alignment is a way of arranging the DNA, RNA or Protein sequences for the identification of similar regions that may lead to the functional, structural or evolutionary relation ship between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.



**Figure 1.1** Sequence alignment of two human zinc finger proteins, Single capital letters = amino acids (a.a), Red = small, hydrophobic and aromatic a.a except Y, Blue = a.a with acidic side chain, Magenta = a.a with basic side chains, Green = a.a with hydroxyl, amine, amide and basic side chains, Gray = others, \* = identical a.a, : = conserved substitutions (same color group), . = semi-conserved substitution (similar shapes), image produced by using ClustalW {10}.

Figure 1.1 shows a simple sequence alignment, produced by Clustal.W {10}, of two human zinc finger proteins. The sequence identity is given on the left hand side by Gene Bank accession number. Sequence based searches are frequently in use in the biological field, mainly two types of sequence searches are dominant including the nucleotide sequence and amino acid sequence searching. Searches can be limited by type of organism, e.g. human, animal, bacteria or virus. One of the main online tools used is BLAST (Basic local alignment search tool), which can be accessed at the NCBI (National Center for Biotechnology Information), part of the U.S. National Institutes of Health.

One of the main objectives of sequence searches and alignment is to identify and align related sequences of proteins with the hope to find structure to function similarity provided that their function and/or structure are already known. The identification of the catalytic and conserved residues is made easier by comparison with multiple sequences. There are many online databases available for submitting the sequence of a novel protein and to compare it with thousand of other proteins whose function is already known. There is a proper system of scoring functions available in databases which give the leading hits in descending order by considering different factors like identical amino acids, similar amino acids, conserved amino acids and catalytic amino acids. The sequence based search is a valuable tool in the preliminary studies of a novel protein and assists in suggesting structure and function.

At the first step of analysis, the use of sequence similarity in the characterization of proteins has been found to be a very powerful but sometimes

insufficient method to infer a protein function completely {11}. It has been observed that some structurally closely related enzymes can catalyze totally different type of reaction {12} and conversely completely unrelated enzymes may catalyze overall the same reaction {13}. For instance in case of AHS (amido hydrolase super family), if seen through a sequence perspective protein Tm0936 has got similarity with the large chlorohydrolase and cytosine deaminase subgroup within the AHS {14}. A successful *in silico* study identified the function of protein Tm0936 from *Thermotoga maritima* as an adenosine and thiomethyl adenosine deaminase {2} which on the contrary shows no sequence similarity with adenosine deaminases {14}. It is being stated that approximately 40% of newly sequenced genes have not been assigned a certain function and if they have, the specificity of the function is either incorrect or not the primary function {15}.

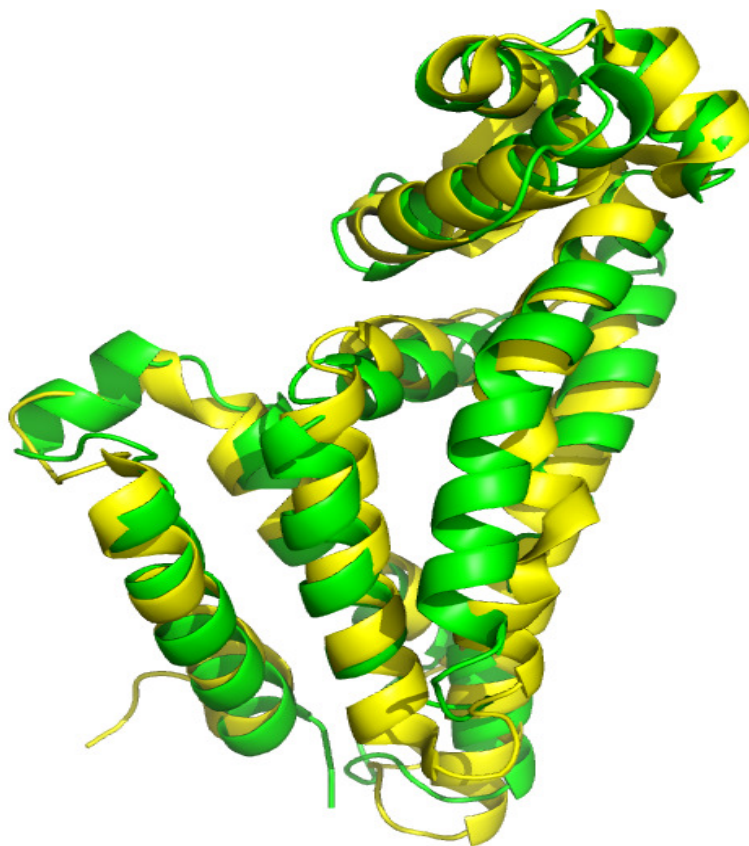
### **1.1.2 Protein Fold similarity**

The amino acid sequence of a protein is directly responsible for the folded structure of the protein. Even with amino acid sequence similarity lower than 20%, it has been observed that the folded protein structure can be still conserved. A number of recent advances have been made for obtaining functional information from protein structure. A fold relationship to an already characterized protein or group of proteins can give some insight into the general function of the uncharacterized protein. Methods of deducing function directly from structure, without the use of evolutionary relationships, are developing rapidly. All such methods can also be used with models of protein structure, but model accuracy imposes certain limitations {16}.

The rapid expansion of the structural genomics field has created a new urgency leading to improved methods for structure-based annotation of function. Structural alignments, which are usually specific to protein and sometimes RNA sequences, are used to get information about the secondary and tertiary structure of the protein {17}. These methods depend on the availability of structural information; they can only be used for sequences whose corresponding structures are known (usually through X-ray crystallography or NMR spectroscopy). As both protein and RNA structure are more conserved in evolution therefore sequence structural alignments can be more reliable between sequences which have >50% sequence homology but for those which

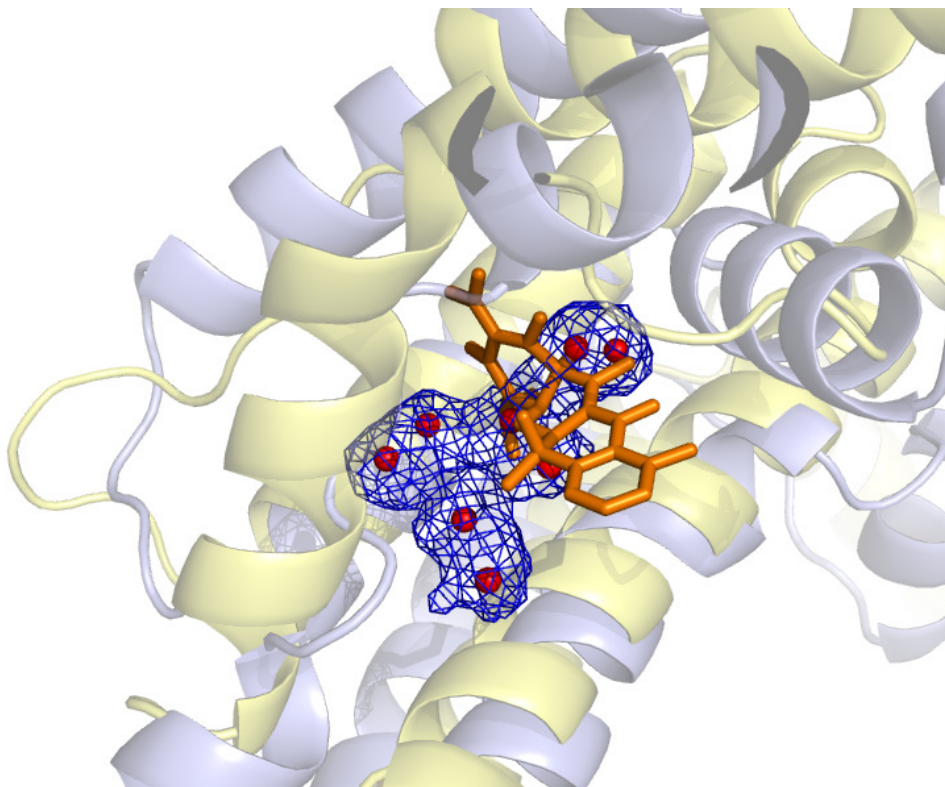
have diverged so extensively that sequence homology drops to 20%, the comparison cannot reliably detect their similarity. Structural alignments are used as the "gold standard" in evaluating alignments for homology-based protein structure prediction. They explicitly align regions of the protein sequence that are structurally similar rather than relying exclusively on sequence information {18}. A large number of structures deposited at the PDB {19} by various structural genomics initiatives {20} are of 'hypothetical proteins', i.e. proteins of unknown function. These proteins are classified as hypothetical when sequence search methods have failed to match them to those proteins that have been functionally characterized. However, if the 3D structure of a protein is known it opens up a possibility of a chance for annotating its function by means of its structural analysis {21}.

Different databases like DALI (Distance mAtrix ALignment) {22} and SSAP (sequential structure alignment program) {23} are being used for the structural and fold alignment purposes. In each program different properties related to similarity like similar contact region, fragment similarity and atom to atom vector similarity are used. For instance crystal structure of a protein (PDB: 3GZI) belonging to TetR family from *Shewanella Sp. PV-4* was submitted and annotated as a transcriptional regulator protein. A comparison of 3GZI by protein fold using DALI search database gave many hits like PDB entries 3DEW, 1T33, 1BHQ, 2D6Y and 2ID6. The best hit (3DEW) is shown together with 3GZI in figure .1.2



**Figure 1.2** Superposition of 3GZI (green) and 3DEW (yellow), image taken from Topsan {24}

3GZI is an example where the crystal structure fortuitously contains a ligand which may help identify the function of the protein by means of sequence similarity search. The structure of 3GZI had unidentified electron density which when compared with PDB: 2TRT (a tetracycline binding protein) corresponded to the tetracycline binding site, though the ligand in the binding pocket is clearly different (Figure.1.3). In a co-crystallized structure it depends on the source of the ligand, if it is an artifact of crystallization then it is less helpful but if the compound has been co-purified with the protein then it represents the native ligand. There are numerous examples where co-factors such as NADP, FMN are co-purified with an unknown protein thus providing an excellent start in assignment of function.



**Figure 1.3** Superposition of 3GZI (light blue) with TerR family protein 2TRT (light yellow), the electron density for unidentified ligand (blue mesh) for 3GZI is located in the same region where tetracycline (orange stick model) is residing, image taken from Topsisn [24]

Structure based functional characterization is used and has had significant successes, but in many cases the results are not valid and some times appear to be incorrect [25]. Identification of a protein's "real" substrate when it has a remarkable substrate diversity remains a challenge [11]. For enzymes of unknown function (uncharacterized), prediction of potential substrates on the basis of structural complementarity is principally an alternative to bioinformatics based functional identification [15, 26]. In terms of possible protein structure along with structural genomic study the structural genomics initiative has generated many protein structures every year [27-30] the function of some of these proteins is known, but for many of them only the 3D structure is known which may help in understanding their biochemical roles [11]. Different amino acid sequences are able to generate proteins with similar folds but this does not always leads to a related function [31]. Structure-based prediction becomes much attractive when the target unknown protein has a weak similarity to proteins with known activity and this makes sequence based methods unreliable [32]. The functional characterization of enzymes encoded

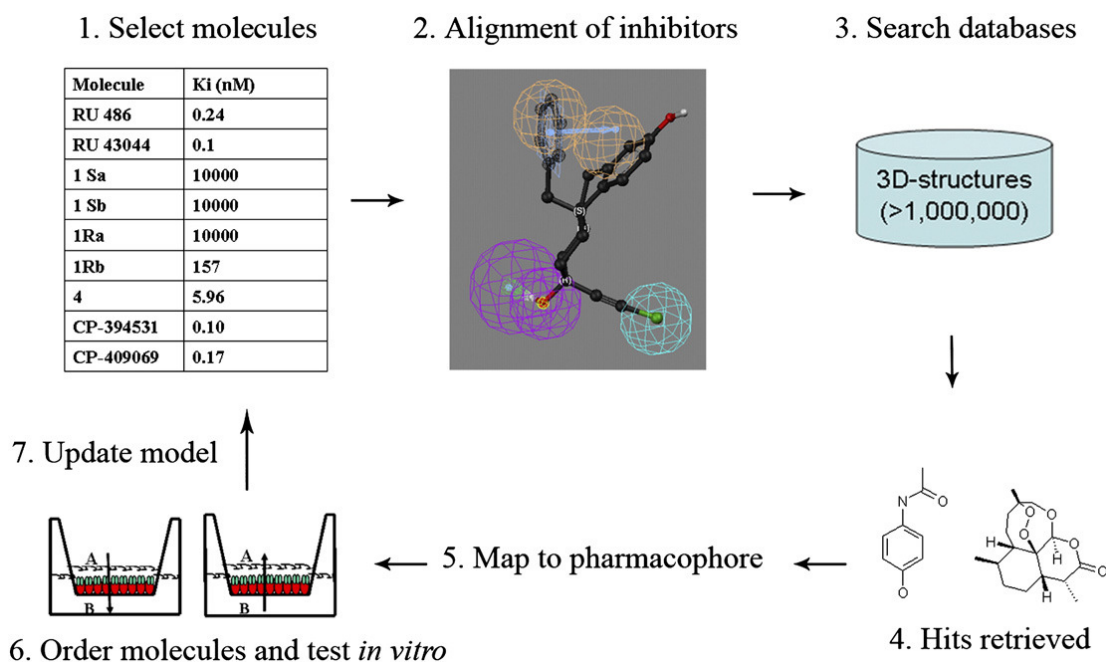


within completely sequenced genomes is painstaking work and requires much dedication {14}.

### 1.1.3 Protein functional Assignment

The number of protein structures determined in an attempt to map the “Protein fold universe” is increasing at a significance pace. Despite the claim that protein structure is intimately related to function. The ability to read out the function of a protein based on its 3D structure is far from simple. Even when the function of a protein has been determined, correct identification of the substrate or co-factor binding mode within the active site can prove illusive.

Normally traditional structure based drug discovery involving *in silico* screening of compounds is carried out by using large multi-conformational libraries of drug like compounds against the target protein crystal structure. The objective is the identification of promising lead compounds (Figure 1.4). The leads can then be modified in iterative rounds with library creation of derived compounds. This is followed by *invitro* screening and structure determination of enzyme-inhibitor key complexes which ultimately leads to the design of new compounds {33}.



**Figure 1.4 Schematic chart for pharmacophore-database searching for structure based drug discovery, image taken from {33}**

More recently to probe the enzyme binding site, fragment methods have been used. The method involves the screening of small fragment libraries and identifies moieties that can be joined together or grown to form lead compounds {11}. Structure-based prediction of enzymatic capabilities is now an important issue, and many researchers worldwide are working, with the ultimate target of providing a suitable *In silico* protocol that, given an experimental structure, can predict its function to some reliable levels {11}. There have been extensive advances in computational field to understand the ligand binding mainly for the detection of the protein functional site. The characterization remains the target of structural bioinformatics, for this purpose various productive techniques have been tested in recent times which are expected to be helpful in both the detection of functional site and for the design of new pharmaceuticals. It is apparent that as the nursery of protein structure information flourishes, newly launched computational tools would be used for the functional characterization and structure based drug design purposes {6}. Due to the lack of complete in-depth knowledge of the metabolic pathways, the dire need of new techniques for the functional characterization of enzymes regarding their mysterious substrates has been sensed, for which different approaches including the computational one's are being launched {34}. To cope with this, technique like pharmacophore based searching can prove to be effective and result oriented from the functional view point. It can be used as a significant tool for the identification of both known and unknown substrates leading to the functional annotation of novel proteins.

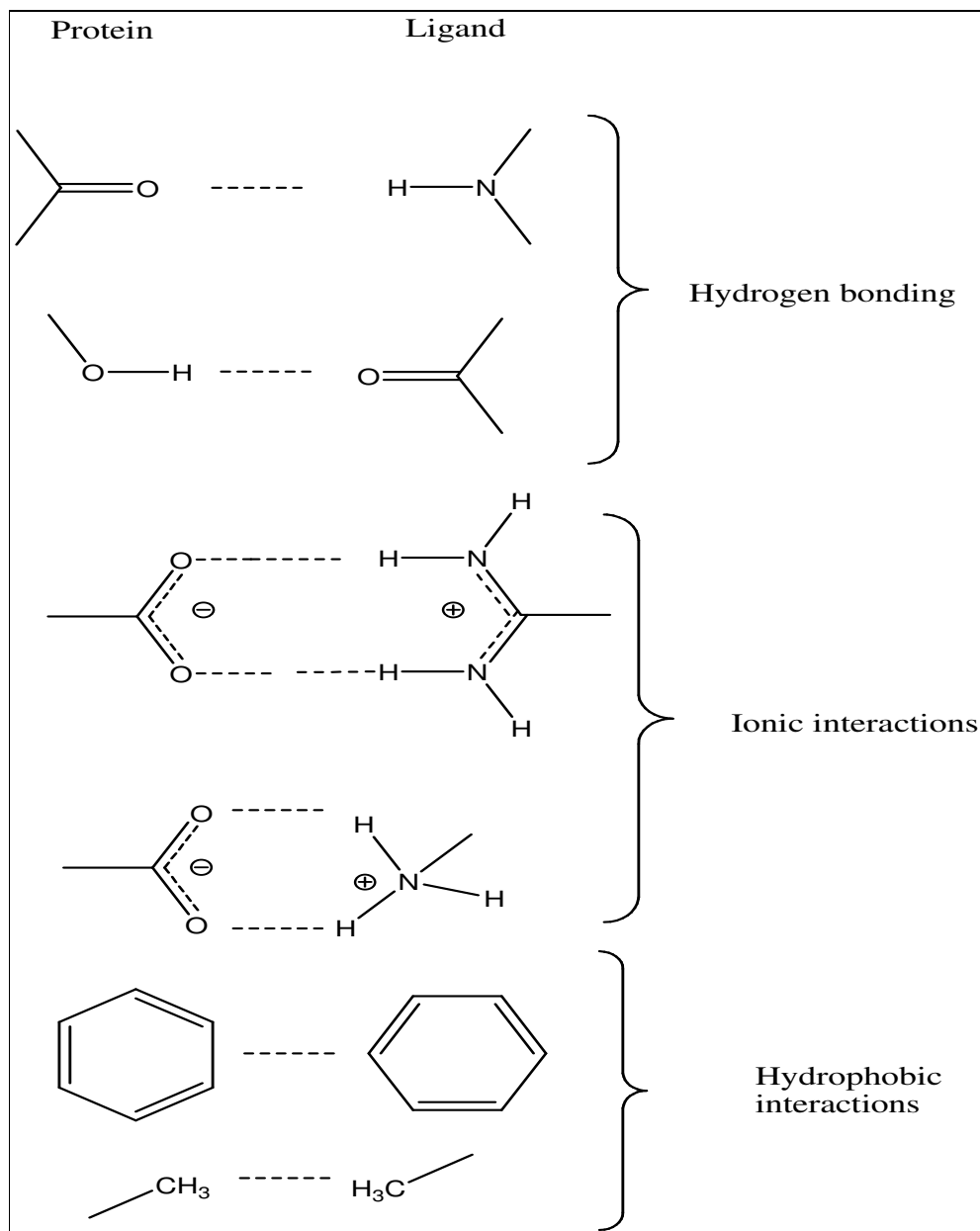
#### **1.1.4 Protein-ligand binding Interactions**

The protein and ligand are fully solvated prior to binding and do not interact. However in the bound state both are partially desolvated and form favorable interactions with each other {35}. Mostly protein-ligand interactions occur as a result of relatively weak forces in the form of electrostatic attractions, such as ion-ion, ion-dipole, dipole-dipole (hydrogen bonding), dipole-induced dipole, London or Dispersion forces and hydrophobic effect. The hydrophobic effect has been studied extensively and is important in driving protein folding and protein ligand binding. Many theories and models have been put forward to describe the hydrophobic effect and in order to support these models and theories various experiments have been carried out. A recent study has supported the hypothesis

that structured water molecules including both water molecules displaced by the ligands and those reorganized upon ligand binding determine the thermodynamics of binding of ligands to the active site of the protein. Despite extensive studies carried out in order to understand the hydrophobic effect, their molecular mechanism remains controversial, and there are still no reliable models for predicting its role in protein ligand binding. The hydrophobic effect is now understood to be a complicated phenomenon, undergoing entropy, enthalpy and free energy changes during protein ligand interactions {36, 37}. The structure of binding site of the protein has crucial importance in specifying potential interaction such as Hydrogen bond donor, Hydrogen bond acceptor and hydrophobic patches. These interaction sites are referred to as “hot spots”. For exploring the binding site of the protein, primarily the amino acids forming the cavity of a given protein describe the recognition pattern. The binding event takes place when the ligand complements to these amino acid residues in the three dimensional space in the binding cavity {38}. There is a large amount of 3D structural data available for protein ligand complexes and their corresponding affinities. Some of the typical features found in all of these complexes are

1. High degree of steric complementarity between the protein and ligand
2. High level of agreement at the ligand protein interface regarding the lipophilic and hydrophilic interactions
3. Ligand binding in an energetically favorable conformation

Typical non-covalent interactions found in protein-ligand complexes include hydrogen bonding, ionic interactions and (lipophilic) hydrophobic interactions. The lipophilic interactions are essentially between the apolar regions of protein and ligand. The water molecules in the lipophilic region of the binding site are unable to form hydrogen bonds with the protein and so may form an ordered network of waters within the binding site. The displacement of these ordered water molecules is generally accepted as contributing to the strength of lipophilic interactions {39}. Figure 1.5 represents some of the common non-covalent interactions present between protein and ligand.

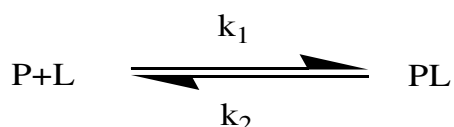


**Figure 1.5 Common non-bonding interactions between the ligand and the protein usually in case of hydrophobic interactions the lipophilic part of protein is in contact with the lipophilic part of the ligand**

In the most well known cases hydrogen bonding phenomenon takes place when an electronegative donor group (e.g. O, N, S) with the capability to withdraw electrons from a proton in a D-H covalent bond ( D= Hydrogen bond donor group), leaving a net partial positive charge on the partly deshielded proton and thus creating an opportunity to form a electrostatic interaction between the partially deshielded proton and another electronegative group. Hydrogen atoms or protons are found almost between every pair of non-covalently bonded heavy

atoms in biological systems, they constitute about one half of the atoms in macromolecules and two-thirds of the atoms in solvating water. Apart from enzyme catalysis, hydrogen bonding plays a very important role in the protein-ligand binding event and is crucial to the recognition of ligands by the protein. The presence of hydrogen bonding certainly stabilizes the overall structure of the protein and influences the binding pose of the ligand in the active site by enhancing the undergoing favorable interactions between the two {40}.

Majority of the protein ligand interactions are reversible in nature. The general equation used for the bimolecular interaction between a ligand and a protein is given as



Where P = free form of protein and L = free form of ligand, PL = protein-ligand complex,  $k_1$  is the association rate (on-rate) and  $k_2$  is the dissociation rate (off-rate) for the protein-ligand complex. The association constant  $K_a$  for the ligand binding is  $K_a = k_1/k_2$ , while the dissociation constant,  $K_d$  is given as  $K_d = 1/K_a = k_2/k_1$ .  $K_d$  value for protein-ligand reaction represents the protein binding sites to be half-saturated with the ligand, and is very useful in identifying the binding affinity, a nano-molar or low micro-molar value of  $K_d$  represents tight binding between the ligand and the protein, while a high micromolar or millimolar value of  $K_d$  represents weak binding {41, 42}.

### **1.1.5 Solvent molecules and protein binding site**

In addition to the desolvation effects during binding of ligand to a protein, certain solvent molecules are important due to potential protein-solvent-ligand interactions. The interactions are mainly based on H-bond donor and acceptor functionalities. These water molecules are identified in moderate to high resolution X-ray structures but importantly not in NMR structures which potentially limits the utility of these structures for ligand binding predictions. Bound water molecules perform pivotal roles in the binding site of the protein; it has been shown that they are involved in mediating the protein-ligand

interactions. The bound water molecules in the binding pocket of the protein have crucial importance in the course of generation and optimization of protein structure based pharmacophore model. More the position of water molecule verified higher would be the contribution of the corresponding pharmacophore feature(s). If water molecules are not well defined, computer simulations may help in identifying the most important water molecules in the binding site, as there are many other factors like geometry of the binding pocket, charge distribution, polarity and ligand binding mode which can influence and alter the results. The responsibility of dealing with these water molecules largely stays in the hands of a modeler to decide their inclusion or exclusion in the binding site {43}.

## **1.2 *In silico* Techniques for functional annotation of Novel proteins**

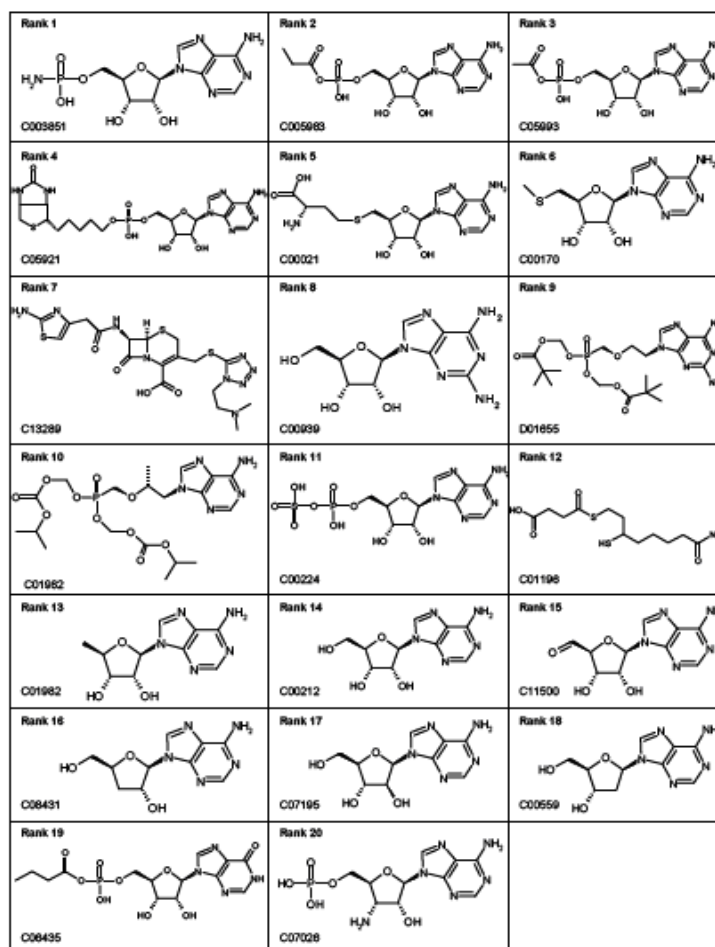
### **1.2.1 Docking**

The docking methodology involves sampling, scoring and ranking multiple conformations and orientations (rotations and translations) of a ligand within the binding site of the protein. In the field of molecular modelling, docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex {44}. The availability of relatively simple to use Docking software such as GOLD {45} Autodock {46} and FlexX {47} are a few to name. Structural biologists now attempt to dock known ligands or screen libraries of compounds against newly solved structures from functional point of view. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using for example scoring functions.

In majority of the docking methods, the search is for the electronic and steric complementarity between the potential ligands and the protein binding site {48}. Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to predict the affinity and activity of the small molecule. Hence docking plays an important role in the rational design of drugs {49}. Given the biological and pharmaceutical significance of molecular docking, considerable efforts have been directed

towards improving the methods used, such as speed and validity of scoring functions. The docking experiments are widely in use but with only limited success. A notable success story is that of the X-ray crystal structure of Tm0936 and the identification of the potential substrates for this deaminase. In this study the X-ray structure of the enzyme had been determined as part of a broad structural genomics effort (PDB codes 1p1m and 1j6p), which can be assigned to the AHS by fold classification and due to the identity of certain active site residues. By sequence similarity, Tm0936 most resembles the large chlorohydrolase and cytosine deaminase subgroup, which is often used to annotate amidohydrolases of unknown function. 14 cytosine derivatives were tested as Tm0936 substrates and no turnover was detected for any of them. In an effort to find the true substrate, database of high-energy intermediates were docked into the structure of Tm0936.

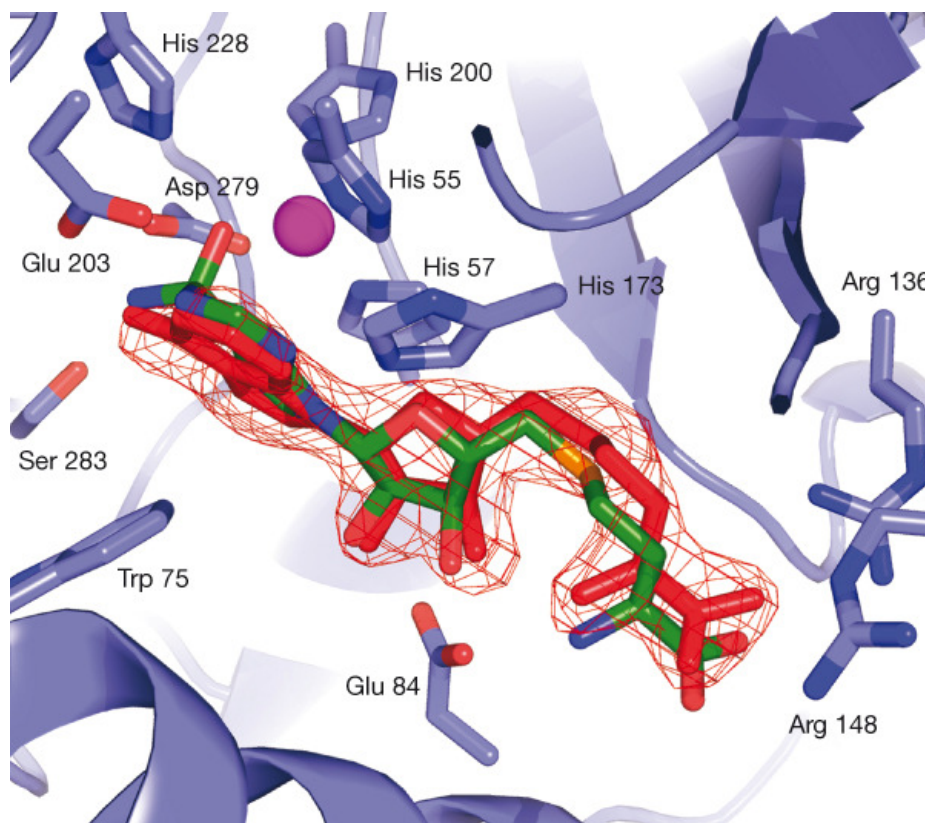
Overall, 28 amidohydrolase reactions operating on 19 functional groups were modelled by these high-energy structures for each of the 4,207 metabolites that bore them, were converted into high-energy intermediate geometries (transition-state-like geometries), with their appropriate charge distributions, leading to the calculation of about 22,500 different forms of the metabolites. Modeling substrates as high energy intermediates, was particularly useful when docking to apo structures in these studies. The molecules best-ranked computationally were dominated by adenine and adenosine analogues. Four of these, including 5-methylthioadenosine and S-adenosylhomocysteine (SAH), were tested as substrates, and three had substantial catalytic rate constants. The enzyme has no obvious sequence similarity to any known adenosine deaminase and the study shows that it exploits interactions not previously identified in the active sites of these enzymes {2}. From the results it became clear that adenine nucleotide compounds were the best hits (Figure 2.1).



**Figure 1.6** Top scoring hits from docking against Tm0936 with a large number of adenosine analogues, image taken from {2}

The five of the top hits were assayed for biological activity and it was found that S-adenosyl-L-homocysteine (SAH) and 5-methyl-thioadenosine were both good substrates with  $K_{cat}/K_m$  values of  $5.8 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$  and  $1.4 \times 10^5 \text{ M}^{-1}\text{s}^{-1}$  respectively. Using SAH the crystal structure of the complex was determined to confirm the mode of binding (Figure 2.2)





**Figure 1.7 Comparison of docking prediction and the crystallographic results. Enzyme carbons are colored light blue, SIH is colored red while carbons of intermediate of SAH (docked) are colored green for differentiation purposes. The purple sphere represents the divalent metal ion, image adapted from {2}**

Figure 2.2 shows superposition of the crystal structure of Tm0936 in complex with SIH (S- inosyl homocysteine) (red) and the docking predicted structure of the high-energy intermediate of S-adenosyl-L-homocysteine (SAH) (carbons in green). It was observed that the docking prediction and the actual structure of the complex had good agreement. On the basis of docking and enzyme activity results, Tm0936 was reported to be an adenosine deaminase [2].

In another study, substrate docking was applied for the prediction of function of one of the member of enolase superfamily. In this study a homology model of the test enzyme was used as a template for docking in place of a crystal structure, and the studies identified the enzyme as an N-succinyl arginine/lysine racemase {50}. It is suggested that Docking is likely to be useful for the general analysis of the newly solved structures as it can speedily identify potential ligands and can lead to the structural functional relationship between the new structure and those already in the PDB {11}. Recently a group has introduced the Bayesian approach by using Gaussian mixtures for the distribution of particular

protein atoms around ligand fragments, which showed that it can lessen the search for the position and type of protein atoms involved with specific ligand fragments, and this technique can lead to the prediction of the type of ligand binding site without having structural information about the protein {51}. Another group has developed a docking methodology that allows the incorporation of pharmacophore type constraints and under these constraints the procedure gives solutions, showing that in docking procedure the pharmacophore-based filters can be incorporated {52}. In some cases the diversified catalytic action of the enzyme in the course of metabolic pathways could be understood provided the functional aspects of the newly sequenced genes are considered {14}. The hypothesis that only a number of selected representatives from large databases of ligands are sufficient to obtain reasonably accurate information regarding the preferred type of substrates for a given enzyme or to put in a general manner the preferred binders for a protein receptor is worth testing {11}.

Docking studies face certain challenges like the choice of database which is used. For a particular enzyme the database may not be having the right substrate in it. Another problem is if the enzyme undergoes conformational changes during catalysis, as a result the conformation of the structure used may not be suitable for docking. This and in many cases the weakness of scoring functions to discriminate between a successful hit and a false hit can produce misleading results {11}. Docking of 1,000,000s compounds is computer intensive and the subsequent scoring and ranking of hits is highly subjective. To reduce the number of compounds to be looked at in detail, pre-screening of compounds can be performed. Docking has been used extensively for finding out the potential ligands (drug candidates) and for detection of the correct orientation of a ligand in the protein active site. In comparison to docking pharmacophore searching seems to be a better technique as the computational cost is much less and the technique is much quicker relative to docking. Despite extensive use, both the techniques have potential draw backs which need attention when dealing with individual cases. Table 2.1 gives a comparative account of limitations associated with docking and pharmacophore searching {48}.

S.No	Docking	Pharmacophore Searching
1.	Assuming the protein binding site to be rigid and the ligand flexible, leading to selection of compounds without visualization of their manual interaction	The static model is based on the single conformation of the active site while in reality the protein structure is very dynamic and thus some of the constraints introduced on the basis of rigid model may not be present and thus the hypothesis may become wrong about some ligands, leading to false positive and non-binders
2.	Without any information about true actives or known experimental complex structures	In majority cases in the 3D structure of the protein the average number of key interactions is typically limited to 4-6 in the form of geometric constraints, while leaving dozens of other significant interactions.
3.	Inability to account for role of bound water molecules	Queries containing large number of exclusion spheres, take very long to search the database.
4.	Inability to Incorporate the conformational flexibility of the ligand and the protein target site	The majority of the information contained in the protein structure is not utilized during the database search, and the result just in the form of “hit” (matching the query) and “no-hit” (not matching the query).
5.	Inefficient sampling of orientational space	Any steric or electronic constraints imposed by the target but not defined in the query are totally ignored
6.	Imprecise and unreliable nature of scoring functions	There is no scoring function

**Table 1.1 Potential drawbacks associated with docking and pharmacophore searching {48}**

### 1.2.2 Pharmacophore

Pharmacophores or pharmacophore, a word which appears simple in meaning and concept, entails a vast amount of knowledge regarding bioactive molecules and their structure-activity relationship. The origins of pharmacophore dates back to 1900 when the term pharmacophore was first used by Paul Ehrlich {53} more than a century ago, describing it as

***“A molecular framework that carries (phoros) the essential features responsible for a drug’s (pharmakon) biological activity”***

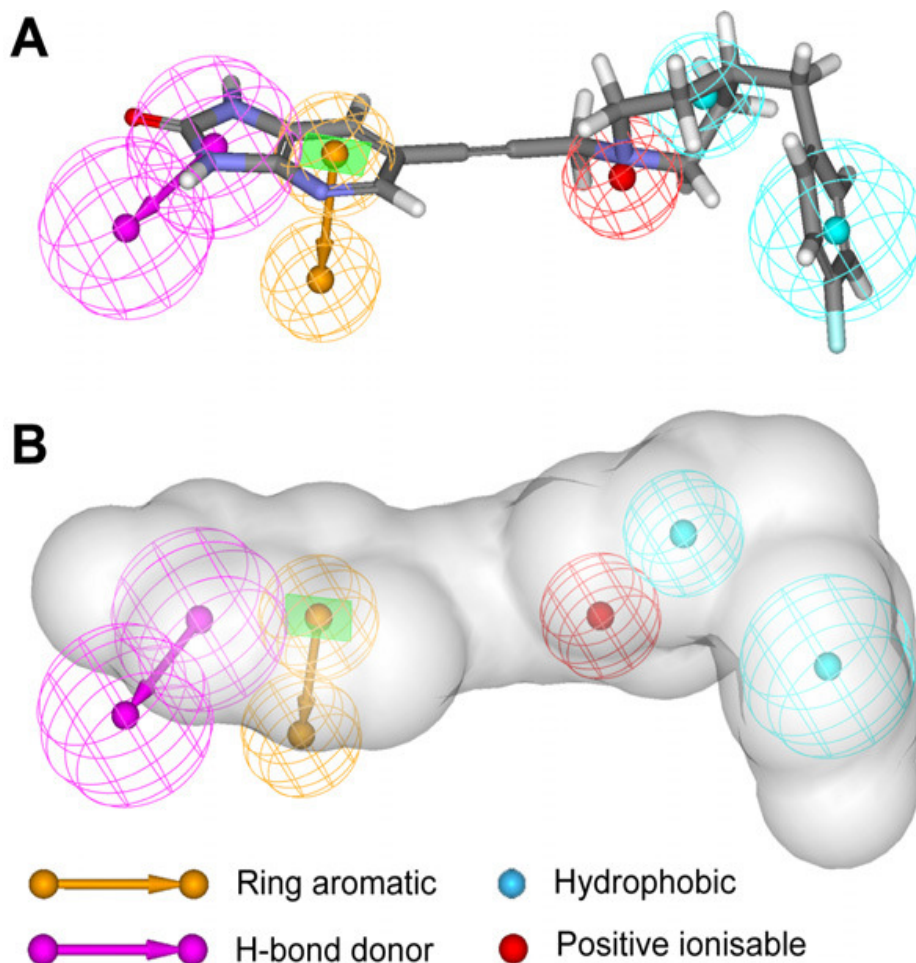
Later on in 1970 Peter Gund {54} defined the pharmacophore as:

***“A pharmacophore is an interpretation of a collection of chemical features located in the 3D space that accounts for the binding of ligands to a common receptor”***

According to the IUPAC recommendations {55}:

***“A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response”***

It has also been defined as “In the event of binding of ligand to its biological target (macromolecule), complementary groups on the protein binding site recognize key features of the ligands and the 3 dimensional arrangements of these features is commonly referred to as a pharmacophore”. Alternatively pharmacophore is described as “specific part(s) of a molecule that confers a molecules biological activity”. More loosely, a compound possessing the required biological activity, albeit without sufficient potency or freedom from unwanted other actions. A lead compound having this activity can be subjected to systematic chemical manipulation to increase its potency, and removing the unwanted actions {56}. Figure 2.3 represents a typical pharmacophore model including different features like H-bond donor, Hydrophobic, positive ionisable and ring aromatic. The hit satisfies all the required features of the pharmacophore and thus fits in well to the model.



**Figure 1.8** Pharmacophore model for NMDAR antagonists. (A) The antagonist molecule satisfying the Pharmacophore model. (B) The van der Waals volume occupied by the antagonist molecule along with the pharmacophore features. The spheres represent the tolerances for various features, image taken from {57}.

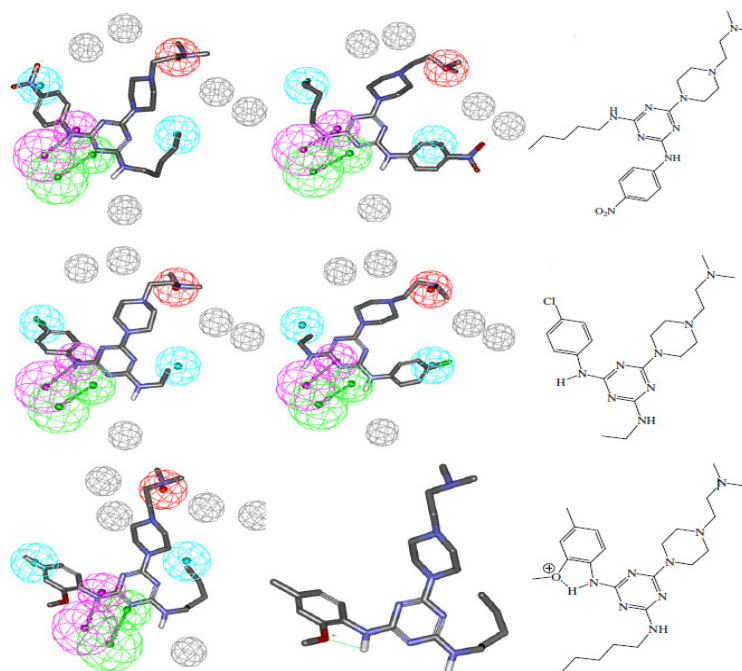
### 1.2.3 Pharmacophore Searching

Pharmacophore models have been used to explore the features associated with protein-ligand interactions for recognition purposes in the binding site of the target protein {58}. Multiple pharmacophores have been used previously and merged together in order to promptly search the database for the removal of unsuitable compounds or to suggest new compounds that can be tested against the binding site of the target protein {59}. Pharmacophore searches have been overlooked in some ways in the last decade, though some of its features have been utilized by the scientific community. However in the pharmaceutical industry the application of pharmacophores is widely used in the drug discovery process, which clearly exhibits the power of this technique {60}.

Predominantly the pharmacophore methods can be classified into two major categories i.e.

1. Protein structure based pharmacophore modeling
2. Ligand-based pharmacophore modeling

Protein structure based modeling which is a direct approach, majority of the features/constraints are introduced based on information available from the 3D structure of the protein binding site and has been widely used as an alternative tool or complimentary to high-throughput screening (HTS). It is considered a method to produce a small more focused library of compounds for screening (hundreds of compounds) rather than the 10's of thousands to millions of compounds that are often screened in a blind HTS. Protein structure based modeling provides added advantage of detail analysis of the binding pocket of the target protein. In majority cases the residues in the binding site dictate the shape and chemical features of the prospective ligands to fit in [38]. In case of indirect ligand-based models (Figure 2.4) the pharmacophores are designed as per chemical features of known active compounds, in which chances to judge and predict both the water-mediated interactions and displacement of bound water molecules in the binding site of the protein is very limited [43].



**Figure 1.9** Ligand-based pharmacophore modeling: novel inhibitors of the Ca<sup>2+</sup>/calmodulin-dependent protein kinase II inhibitors with alternative mappings against the pharmacophoric features, image adapted from [61]

The general problem of finding compounds/ligands which can successfully bind to a receptor is a challenging one. More often it is believed that if the 3D structure of a given protein is known, then the route to find suitable ligands is comparatively easy. From the 3D structures of ligand bound or unbound proteins, it becomes possible to pinpoint the key residues of the protein involved in ligand interactions. Using this knowledge a pharmacophore can be generated to screen for compounds which share the common features of the pharmacophore. In brief the availability of the 3D structure of a protein active site offers significant information which is crucial in the development of pharmacophore models. However in practice, to find a ligand which can satisfy all the requirements in relation to the complex features in the active site of the protein is quite complicated as it is not only a matter of structural and computational biology but also of chemistry {48}. In pharmacophore searching it has been shown that the essential features contributing towards ligand binding are both from the protein and ligand. In relation to the pharmacophore modeling the use of multiple excluded volumes is extensively used {62, 63} to supplement the stringency conditions on the model and to selectively describe the shape of the binding pocket. The purpose of introducing exclusion spheres is to avoid steric clashes, and the penetration of ligand structures into the space occupied by the amino acid residues in the binding site. It has been found that the inclusion of exclusion spheres in the pharmacophore models improves the pruning of databases so that large multi-conformational databases can be searched in 1-2 hours.

In order to be recognized as a useful tool, pharmacophore model has to provide certain information regarding protein ligand interaction. For instance the functional groups directly involved in the interaction, non-covalent bonding and inter charge distances. Different analogies have been used to describe the concept of pharmacophore and its role, an interesting one is Fisher's Lock and Key model. When dealing with the change in conformation of glucoside on its contact with the enzyme, the enzyme and glucoside must fit together like the key in the lock, so that to have a chemical effect on each other. Though it represents the target protein to be a rigid structure but is still in use nowadays. Another representation is of "hand in a glove" which appears more precise as it takes into account the steric complementarity, chirality and flexibility of both protein and ligand {64}.

To make a “pharmacophore” the method relies on the identification of a limited number of hydrogen bond donors, hydrogen bond acceptors ionisable features and hydrophobic groups within the active site. These are the essential interactions that a compound should make with the protein. If the “essential interaction” are not known then normally a series of pharmacophore are constructed and screened {65}. A well-defined pharmacophore model includes both hydrophobic volumes and hydrogen bond donor and acceptor vectors and a shape constraint or excluded volume (which corresponds to the extent of the active site cavity) {66}. By varying the tolerances on each of these interactions the number of potential compounds can be reduced rapidly from several hundred thousand to about 20-100 compounds. The strength of pharmacophore searching is its speed and an ability to interrogate an active site in a systematic way. It is possible to include or exclude ordered water molecules which may or may not be important for ligand binding. The method uses the Accelrys packages, Cerius2 and Catalyst for performing the search. These are the industry standard programs and no other software, free to academic institutions, are available which can permit this kind of search. The aim is not to optimize the use of this set of programs but is rather to use them to validate the application of this method for structure-function prediction.

#### ***1.2.4 Pharmacophore generating programs***

Both protein structure based and ligand based pharmacophores have been created by using several computational programs, such as DS viewer pro which allows the manual generation of pharmacophore {67}, and LigandScout {65} an automated program for creating pharmacophore models from 3D complexes. a significant aspect of LigandScout program is that it considers the water molecules in the protein structure as an integral part of the protein milieu and automatically identifies the protein-ligand interactions mediated through water molecules in the binding site {43}. One of the widely used programs is Catalyst® {68}, within the catalyst program the hydrogen bond donor vector is defined as vector from the donor atom of the ligand towards the corresponding acceptor atom of the binding site of the protein, and vice versa the H-bond acceptor vectors are defined.



In another method called Structure based focusing (SBF), Cerius2 {69} application has been used for protein structure based pharmacophore to identify ligands which can bind to the active site. In SBF method the interaction map of the active site of the protein is generated by using Ludi {70} mainly consisting of H-bond donor, H-bond acceptor and lipophilic features. This results in the creation of a 3D Catalyst query, which is then used to screen the catalyst database to obtain ligand hits that are probable to bind the active site {48}. The method shows that increase in the number of constraints in terms of query features decreases the number of hits. The number of hits decreased to almost half when a 6-query feature pharmacophore was used instead of a 4-query feature pharmacophore. This demonstrates that the increase in number of query features increases the selectivity of the query. It has been shown that the receptor-based pharmacophore appear less biased towards the dataset of compounds. It has been concluded that ligand binding to a receptor is a complex process, where both the ligand and the receptor have to accommodate each other to execute binding {71, 72}. The computational pharmacophore approach is challenging for the researcher as it requires preparedness towards the chemistry, biophysical chemistry, computational methods as well as computational models for analyzing various proteins {73}.

### ***1.2.5 Visualization of Hits***

3D pharmacophore models have a huge potential to be used as a visualization tool. The 3D representation of the pharmacophore features mapped on a protein-ligand complex facilitates the user to intuitively explore the interactions patterns and further to make sensible adjustments {74}. By and large the most direct way by which a pharmacophore model can get validated is when exhibiting over all complementarity and consistency with the interaction pattern specified in the target protein. The use of eye in many cases is the primary and most reasonable judgment to evaluate the pharmacophore hypothesis. First approval is from the expert to rule in or rule out unreasonable solutions prior to final decisions. Visual examination of the pharmacophore models and the hits obtained from these models is of utmost importance. It remains the best approach to fulfill the loop holes often left unaccounted in the computational programs {75}.

### **1.2.6 Hits Screening**

During the course of generation and optimization of pharmacophore models, through virtual screening the database search gives compounds in the form of hits which satisfy all the constraints of the pharmacophore. These compounds will have a very small proportion of biologically active (confirmed through wet lab) and are termed as true positives (TP). A larger portion among the hits includes those compounds which show conformity with respect to the pharmacophore model but are biologically inactive (confirmed experimentally), these are termed as false positives (FP). Those compounds which are not selected as hits from the database and are biologically inactive are called true negatives (TN). There would be certain compounds which are not selected as hits because of non-conformity to the pharmacophore model but are able to cause biological response and are known as false negatives (FN) [60].

### **1.2.7 Future of Pharmacophore Searching**

With the added advantages of computational cost much less than docking and provision of better understanding of the interaction between protein and ligand the prospects of this approach are very bright. Currently there are few if any freely available pharmacophore technologies. This is not helpful at all for the academic groups. The availability of pharmacophore generation models as an open source technology for sharing and application purposes can hugely increase its development. This will certainly expand the range of pharmacophore applications [76].

Experimental approaches to investigate protein ligand interactions

Various techniques are in use nowadays to detect protein ligand binding events. X-ray crystallography emerges as the most preferred technique in revealing the intricacies of protein-ligand binding. NMR with the specificity to even measure the extent of binding and is thus useful to demonstrate sensible information in cases when the ligand binds very weakly to the target protein. Some techniques are aimed for measuring the thermodynamic parameters during binding e.g; differential scanning calorimetry (DSC) and isothermal titration calorimetry (ITC). If the protein is not an enzyme then tight ligand binding can be detected directly by using isothermal titration calorimetry (ITC). Some techniques help to

identify the ligands as binders or non-binders against the target protein e.g. surface Plasmon resonance and chemical micro arrays. Brief account is given below about some of the approaches which are frequently used to investigate protein ligand binding events.

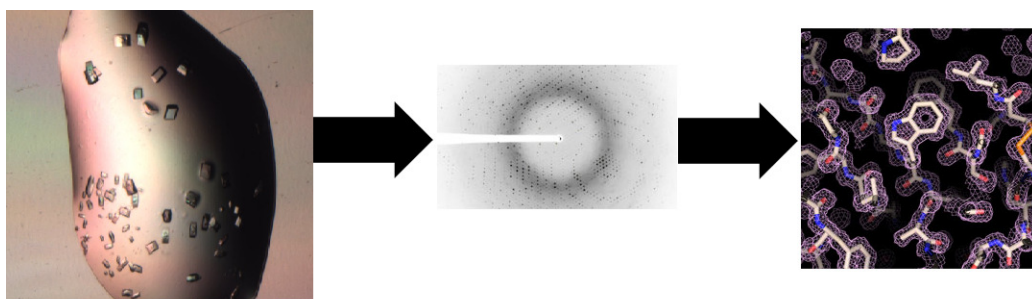
## 1.3 Biophysical techniques for binding studies

### 1.3.1 X-ray Crystallography

The leading work by Perutz and Kendrew in the 1950s on the structure of Haemoglobin and Myoglobin led to a slow and steady start for using X-ray diffraction as a useful technique for protein structure determination. In the last decade protein X-ray crystallography has made rapid progress. It was to more extent triggered by the sudden growth of interest in structural genomics and impressive technological advances further facilitated its development. Now a days it is no longer a tough task to collect the full data set in few minutes and thus solve the structure in hour's time. However the bottleneck remains there which is to grow quality crystals and then get them diffracted at a reasonably high resolution {77}. X-ray crystallography has been successfully utilized for postulating the mechanism of action of enzymatic proteins on their substrates. In one of the case studies, the X-ray crystal structure of *Mycobacterium tuberculosis* shikimate kinase (SK) with bound shikimate and adenosine diphosphate (ADP) was determined to a resolution of 2.15Å. The X-ray structure showed that the binding of shikimate in a shikimate kinase occurs in a specific pocket lined with hydrophobic residues and interacts with several highly conserved charged residues. The comparison with an earlier binary SK-ADP complex showed that conformational changes occur on shikimate binding within the substrate-binding domain. The detail knowledge of shikimate binding proves to be an important step in designing inhibitors for SK. The potential inhibitors can be used as novel anti-tuberculosis agents {78}. On numerous occasions X-ray crystallography has proved to be a vital technique by providing valuable information in understanding the catalytic mechanism of different enzyme {79-81}. Various other ligand binding {82-88} and enzyme inhibition {89-94} studies have been carried out by using X-ray crystallography with an aim to reveal the events taking place at atomic level. 3D structure allows the understanding of biological processes at the most basic level e.g. which molecules interact?, how they interact?, how enzymes catalyze reactions?, and how drugs act? In some

cases, it can help to understand disease at an atomic level, such as the sickling of red blood cells. Protein crystallography can be divided into 3 main steps (Figure 3.1)

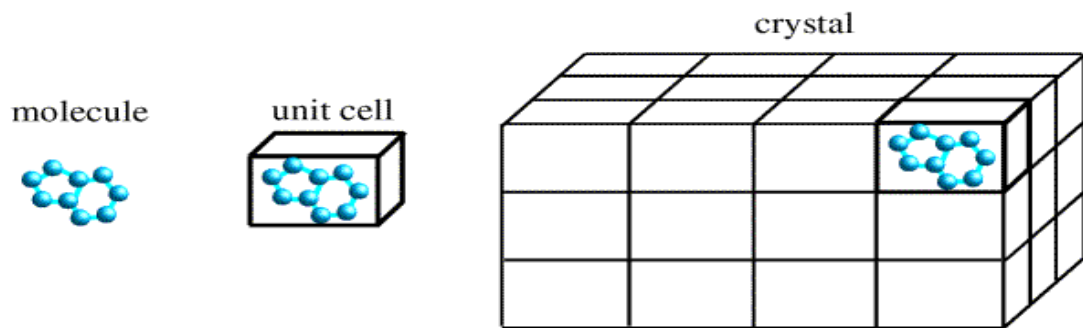
1. Protein crystallization
2. X-ray data collection
3. Structure solution



**Figure 1.10** Schematic diagram for protein crystallization, starting from getting crystals to diffraction images and finally to structure

### 1.3.1.1 Protein crystallization

X-ray scattering from a single molecule would include scattering from air and water and would be incredibly weak and extremely difficult to detect above the noise level. A crystal arranges huge numbers of molecules in the same orientation, so that scattered waves can add up in phase and raise the signal to a measurable level, thus a crystal acting as an amplifier (Figure 3.2).



**Figure 1.11** Comparison of a molecule, unit cell and crystal of a protein with respect to scattering

If a similar protein has already been crystallized then it is definitely worth trying the same conditions to grow crystals of the target protein. In any case if enough protein is available it can be subjected to one or more sparse matrix screens. The methods used normally for obtaining crystals are sitting drop and hanging drop vapor diffusion methods which have the advantage of being the least expensive on protein. In the former method the protein solution is mixed with the precipitant solutions and the protein is slowly brought to super saturation state thereby allowing crystals to form and grow. If the conditions are favorable, at some point during this process the protein becomes supersaturated and is driven out of solution in the form of crystals. Too often these trials result in precipitate or the formation of salt crystals, or nothing happens at all and the drops remain clear. The success rate at this stage is less than 0.1%. If one or more "hits" are obtained in the screens, then the conditions are further optimized e.g. varying the concentrations of all components in the crystallization, slight pH changes, using additives, switching to similar buffers or precipitants in order to achieve large single crystals. Typically it takes few days to several weeks to grow good crystals.

### 1.3.1.2 X-ray data collection

In order to get information about the crystal symmetry, the unit cell parameters, the crystal orientation and the resolution limit. The individual crystal is exposed to an x-ray beam and flash frozen with liquid nitrogen to avoid damage to the crystal. The x-rays are then focused on the crystal and the crystal is rotated through a small angle, typically 1 degree and the X-ray diffraction pattern is recorded.

When x-rays are scattered from a crystal lattice, peaks of scattered intensity are observed which correspond to the Bragg's law:

$$n\lambda = 2d\sin\theta$$

where,

$\lambda$  = the wavelength of the x-ray

$d$  = the spacing of the crystal layers (path difference)

$\theta$  = the incident angle (the angle between incident ray and the scatter plane)

$n$  = an integer

The law states that when the x-ray is incident onto a crystal surface, its angle of incidence ( $\theta$ ) will reflect back with a same angle of scattering ( $\theta$ ) when the path difference ( $d$ ) is equal to an integer ( $n$ ) of wavelength ( $\lambda$ ), a constructive interference will occur. The condition for maximum intensity contained in Bragg's law above allows to calculate details about the crystal structure, or if the crystal structure is known, to determine the wavelength of the x-rays incident upon the crystal.

If the diffraction pattern is very crowded, then the rotation angle is reduced so that each spot can be resolved on the image. This is repeated until the crystal has moved through at least 30 degrees and sometimes as much as 180 degrees depending on the crystal symmetry.

### 1.3.1.3 Structure solution

#### 1.3.1.3.1 *Fourier theory*

The diffraction pattern is related to the diffracting waves through a mathematical operation called the Fourier transform. The electron density is used as a mathematical function, and the diffraction pattern is the Fourier transform of that function. To produce electron density it is necessary to have all the Fourier terms i.e. in the case of single crystal x-ray diffraction this is the full set of indexed reflections  $F_{hkl}$  for the x-ray dataset. However  $F$  is a complex number with both a real and imaginary component. The real part  $|F|_{hkl}$  can be obtained from the integrated intensities from the diffraction spots. The imaginary component, the relative phase angle is lost. This Phase Problem arises as there is no practical way of measuring the relative phase angles for the different diffracted spots experimentally. The numbers of photons give the intensity, which turns out to be proportional to the square of the amplitude (peak height) of the diffracted wave. The relative phase angle for each reflection cannot be measured directly but can be deduced indirectly. There are basically two approaches:

#### **1.3.1.3.2 *Isomorphous replacement and related methods***

If there is no starting model, then Isomorphism Replacement method can be used. In this method one or more heavy atoms are introduced into specific sites within the unit cell without perturbing the crystal lattice. Heavy atoms are electron dense and give rise to measurable differences in the intensities of the spots in the diffraction pattern. By measuring these differences for each reflection, phase angles are derived by using vector summation methods. In practice data from one or more heavy atom derivatives is required to get good enough phases, hence Multiple Isomorphous Replacement (MIR). This method has been extended that heavy atoms can be introduced directly into the protein in the form of selenomethionine. The use of tunable wavelength X-ray sources mean that data can be collected at multiple X-ray wavelengths on the same crystal and the differences can be used to solve the structure.

#### **1.3.1.3.3 *Molecular replacement method***

If the coordinates of a similar protein are available then the structure can be solved by using a process called Molecular Replacement. The method involves rotating and translating the available model into the new crystal system until a good match to experimental data is obtained. The amplitude and phases from this solution can be calculated to see the agreement with the experimental data. If the solution is good enough then the model phases can be combined with experimental data to produce an electron density map.

#### **1.3.1.4 Resolution**

In an ideal case, if the molecules throughout the crystal are in identical conformations and the crystal are perfectly ordered, then all the molecules would scatter in phase regardless of the angle of scattering and diffraction data would be collected limited only by the wavelength of the X-rays. This would result in electron density map with peaks at each of the atomic positions. But in reality Proteins are generally fairly flexible, and crystals have lattice disorder i.e. the repeating units are not necessarily perfectly aligned throughout the crystal. This is why when trying to see the finer details of the structure by going to higher scattering angles; the diffraction pattern starts to cancel out. For this

reason, most protein structures are limited to a level of detail where atoms are not resolved from one another. Thus typically tubes of electron density are seen for atoms that are bonded together.

#### **1.3.1.5 Fitting and refinement**

Early in a structure determination, the phase information is usually poor, so the electron density maps are not ideal and does not resolve individual atoms. As a result, the initial model has a lot of errors. The normal procedure is to fit a protein backbone first then if the resolution permits, insert the sequence. Fitting models to density requires the use of computer graphics programs such as Coot [95]. An atomic model can be improved to a great deal by a process called refinement, in which the atomic model is adjusted to improve the agreement with the measured diffraction data. This will have the effect of improving the phases which results in clearer maps and therefore better models. The improvement of an atomic model is judged through R-factor, this is a measure of the agreement between the model and the data - the lower the value the better the model. R-factor is simply the average fractional error in the calculated amplitude compared to the observed amplitude. A good structure will have an R-factor in the range of 15% to 25%.

#### **1.3.1.6 Validation**

In order to get the observation-to-parameter ratio for protein structures, for each atom there are 3 or 4 parameters (3 describing its position, possibly one indicating how mobile it is). At a typical resolution, there will only be about one observation for each parameter. The diffraction data is supplemented with restraints on geometry, which keep the bond lengths, angles and close contacts in a reasonable range. A valuable way to detect this is to leave out a fraction of the data from use in refinement, which can be used to compute an R-free, which is an unbiased indication of the quality of the structure. The main-chain torsion angles are hard to restrain in refinement but the distribution of these angles in the Ramachandran plot is very restricted. It is a way to visualize backbone dihedral angles  $\psi$  against  $\phi$  of amino acid residues in protein structure. Ramachandran plot shows in theory which values, or conformations, of the  $\psi$  and  $\phi$  angles are possible for an amino-acid residue in a protein. It also shows the



empirical distribution of data points observed in a single structure in usage for structure validation. The Ramachandran plot is thus a good indicator of the quality of a structure. Other indicators used are, the distribution of hydrophobic and hydrophilic amino acids.

By and large X-ray crystallography is the most influential practical method for the in-depth analysis of protein-ligand interactions and also helps in understanding the binding mechanism from the high resolution X-ray structure. The idea of co-crystallizations of proteins with different ligands in order to experimentally determine the preferred locations of binding and binding mode is frequently used and has been successful in designing strong inhibitors for certain enzymes. In cases of enzymatic proteins, with the availability of high resolution structures (1.5Å-2.0Å) the electron density map of the target protein helps directly to identify the amino acid residues involved in binding event and also helps in understanding the mechanism of action. Crystal soaking experiments are also carried out in X-ray crystallography with the aim to find binders and their particular binding modes in the active site of the protein {96}.

### ***1.3.2 Nuclear magnetic resonance (NMR)***

NMR has recently emerged as an outstanding technique to demonstrate the dynamics of protein-ligand binding in terms of weak or strong binding. It is based on the magnetic properties of atomic nuclei and has now acquired huge importance in analyzing the structural dynamics of proteins {97}. Using this method, Fischer and Jardetzky {98} detected and characterized the binding of penicillin to the serum albumin. In NMR, protein-ligand investigations rely mainly on two aspects of binding event; one is the creation of protein-ligand complex, and the other is chemical shift perturbations caused due to protein resonances {99}. NMR spectroscopy plays a very significant role among other biophysical techniques used for the analysis of protein-ligand binding events. Various analyzing approaches either from the ligand or protein perspective have been used in this regard, which commonly share the capability to measure weak binding for specific binders {100}. Normally NMR measurements for such purposes are carried out by using <sup>15</sup>N-labelled protein. Changes in chemical shifts as a result of addition of ligand to the protein, relative to the protein spectra alone demonstrate the binding event. If the chemical shifts assignment has

already been carried out then both the location of the binding site and the ligand affinity can be determined {101}.

### 1.3.2.1 NMR brief overview

Atomic nuclei with an odd number of nucleons (protons or neutrons) can have a magnetic dipole moment ( $\mu_B$ ) associated with their nuclear spin. In a magnetic field ( $B$ ) they act as tiny bar magnets and tend to line up in the direction of the field. Nuclei of particular importance are  $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$  and  $^{31}\text{P}$ . According to quantum mechanics these all nuclei are with a  $\frac{1}{2}$  spin and can exist in either parallel (low energy,  $E = -\mu_B B$ ) or anti-parallel (high energy,  $E = +\mu_B B$ ) state in a magnetic field. The energy separation between these two states is:

$$\Delta E = 2 \mu_B B$$

Where  $B$  is the strength of magnetic field experienced by the nucleus, and  $\mu_B$  is the component of the nuclear magnetic moment along the field axis.

Majority of the nuclei are in lower energy state, but in the presence of applied magnetic field at a certain frequency, the nucleus after absorbing energy from the applied field may switch between the two states (Figure 3.3.)

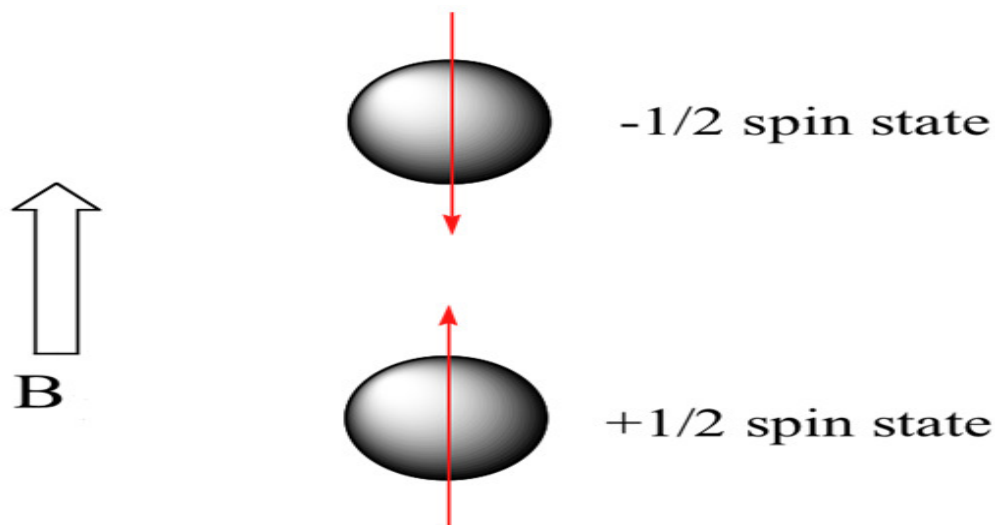
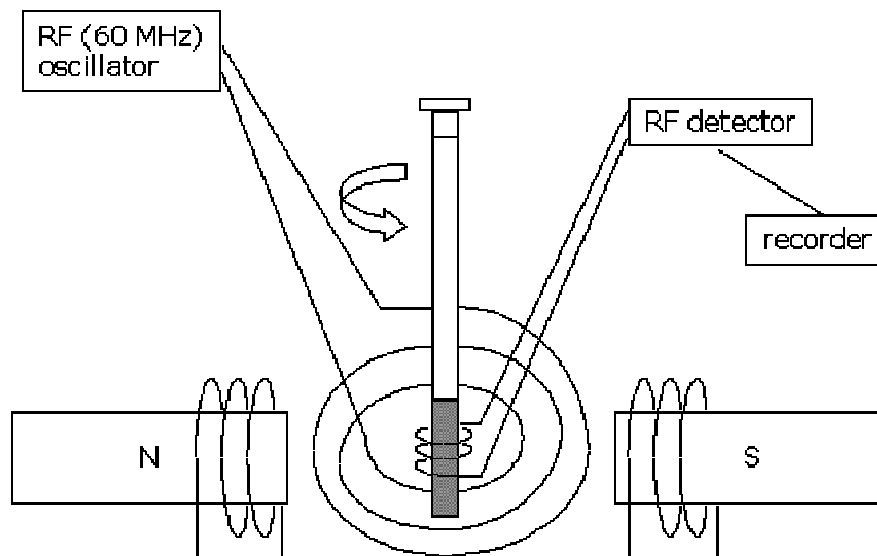


Figure 1.12 Possible spin states of a nucleus in a magnetic field ( $B$ )

In NMR spectroscopy, samples are placed in a strong magnetic field and the resonant frequencies for the nuclei are determined by using an applied radio

frequency (RF) usually in 100MHz to GHz range. The observed resonant frequency ( $\nu$ ) for a particular atomic nucleus is then recorded, which depends on two factors; the nature of the nucleus ( $\mu_B$ ) and the magnetic field ( $B$ ).



**Figure 1.13** Schematic diagram for a typical NMR spectrometer.

$B$  is made up not just of the external applied field from the magnet ( $B_0$ ) but additional contributions from other nuclei and chemical groups. This is because nearby atoms and groups (acting as small magnets) will screen or shift the local magnetic field experienced by the nucleus. Consequently, NMR resonant frequencies are extremely sensitive to the chemical environment of the particular atomic nucleus, and this can be used to give accurate structural information. The chemical shift and spin-spin coupling are ways of describing these additional contributions.

In a modern NMR spectrometer, the Fourier transform approach is used in which the applied magnetic field is kept constant and the magnetic nuclei in the sample are perturbed by applying a short (1-20  $\mu$ s) broad-frequency microwave pulse that excites a range of different nuclei at once. The perturbed nuclear spins relax back to the equilibrium position which is called free induction decay

(FID), which is followed by using field coils or probes placed close to the sample in the NMR spectrometer {97}.

### 1.3.2.2 NMR: Two-Dimensional Spectra

When two atoms are connected by a covalent bond, energy (or more specifically magnetization) transfers between the two nuclei via the bonded electrons. To characterize molecular structures and dynamics by NMR, the first step is to determine the ringing frequency (or chemical shift) for each atom in the protein. This is termed as *chemical shift assignments*. For chemical shift assignments of proteins, NMR spectroscopists almost invariably start with the backbone amide protons ( $^1\text{H}_\text{N}$ )—the hydrogens attached to the nitrogens in the peptide bond. In general, each residue of protein (with the exception of the N-terminal residue and prolines) has one amide proton, so there is approximately one peak for each residue of the protein. In NMR, we add extra dimensions by transferring magnetization between coupled atoms, such as J-coupled nuclei. This is called “two dimensional” NMR, and it involves four basic steps (Figure 3.5)

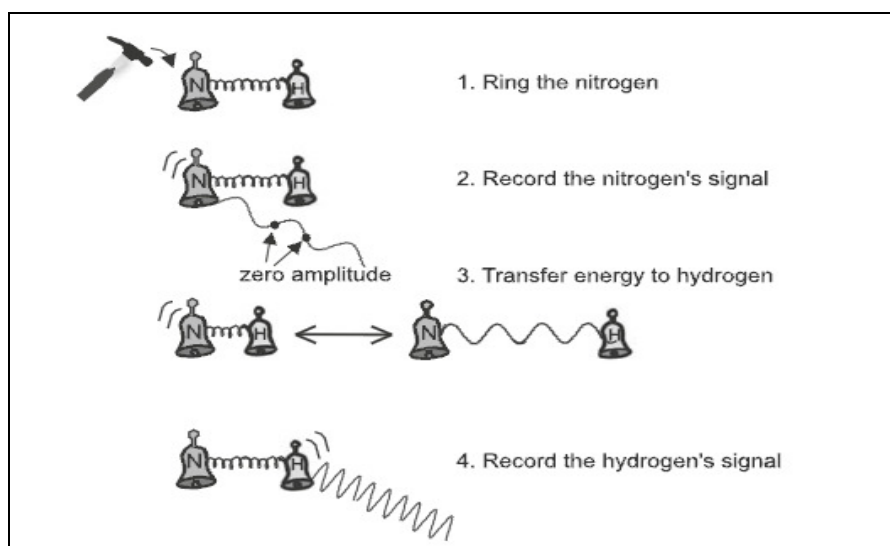


Figure 1.14 Four main steps in  $^1\text{H}$ - $^{15}\text{N}$  HSQC experiment to connect the chemical shift of an amide hydrogen to the chemical shift of covalently attached nitrogen, image adapted from {102}

The result is a two-dimensional map of the NMR spectrum. The ringing frequencies (or chemical shifts) of the amide hydrogens are along the  $x$ -axis and the ringing frequencies for the attached nitrogens are along the  $y$ -axis. In this spectrum, there is a peak for each  $^1\text{H}_\text{N}$  atom bonded to a  $^{15}\text{N}$  atom, or each “ $^1\text{H}_\text{N}$ -

$^{15}\text{N}$  pair.” The location of the peak is at the point  $x$  = hydrogen chemical shift,  $y$  = nitrogen chemical shift. The cross-peak for an amino acid residue can be found on the two-dimensional NMR map if the chemical shifts (or ringing frequencies) for its amide hydrogen and nitrogen are known. By drawing horizontal and vertical lines from the respective chemical shift values, the residue is where the two lines intersect. The most important spectrum in biomolecular NMR is the HSQC spectra.

### 1.3.2.3 HSQC spectra

The  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectrum shows which protons and nitrogen atoms are connected by a single covalent bond. The spectrum in Figure 3.6 A is called a  $^1\text{H}$ - $^{15}\text{N}$  *Heteronuclear Single-Quantum Coherence* correlation spectrum, or HSQC. Here are the five most important attributes of the  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectrum

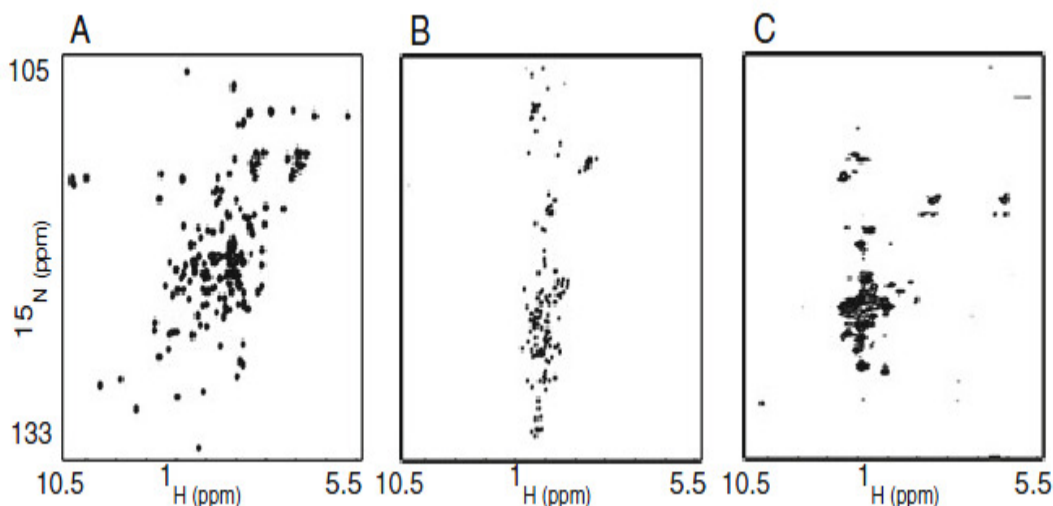
**(a) The Axes:** The  $x$ -axis gives the ringing frequency or chemical shift of each protons attached to a nitrogen atom. The  $y$ -axis gives the ringing frequency of each nitrogen atom attached to a proton.

**(b) The Cross-peaks:** A cross-peak appears in the center of the spectrum wherever a hydrogen is bonded to the nitrogen.

**(c) Total Number of Peaks:** The total number of peaks is approximately equal to the total number of residues in the protein. If the spectrum has more peaks than this, the protein is probably adopting multiple conformations or forming heterogeneous oligomers (i.e., heterodimer, heterotrimer, etc.). If the spectrum has fewer peaks than predicted, regions of the protein are probably dynamic, exist in multiple confirmations, or partially unfolded.

**(d) Peak Intensity and Shape:** For “well-behaved” proteins, all peaks should have approximately the same line width and therefore peak height, like the spectrum in Figure 3.6 A. If a significant number of peaks are weak and broad (or completely missing), regions of the protein are probably dynamic and/or exist in multiple conformations.

**(e) The Peak Pattern:** Peaks should be relatively well separated with minimal overlap (Figure 3.6 A). This is a good sign that the protein is well folded and not aggregating. If peaks are sharp but largely grouped together between  $\delta = 7.5$  and 8.5 ppm the protein is probably unfolded in a random coil-like configuration (Figure 3.6 B). If the peaks coalesce into one big blob near the center of the spectrum, the protein is probably aggregating or forming high-order oligomers (Figure 3.6 C).



**Figure 1.15 Representation of <sup>1</sup>H- <sup>15</sup>N HSQC spectra for (A) a protein properly folded into one configuration, (B) a completely unfolded protein, and (C) an aggregated protein in multiple configurations, adapted from {102}**

The <sup>1</sup>H-<sup>15</sup>N HSQC provides substantial information about the state of the protein even before assigning the chemical shifts to specific atoms.

#### 1.3.2.4 Chemical Environment and chemical shift

NMR has great utility in studying macromolecules and the undergoing interactions. Each atom in the protein resonates at a different frequency. This is due to the fact that each nucleus in a molecule is surrounded by electrons and on top of that each atom exists in slightly different chemical environment relative to each other. This is the reason that if there is some binding event taking place inside the protein, the result is seen in the form of change in chemical shift values for those particular atoms whose chemical environment has changed due to the interaction {102}.

### 1.3.2.5 NMR sampling

A typical protein NMR comprises of the following steps

- 1) Sample preparation, sample concentration in the range of 0.2–2.0mM (several mg/mL) is usually required and samples need to be enriched in isotopic forms of nuclei ( $^{13}\text{C}$ ,  $^{15}\text{N}$ )
- 2) Data acquisition and processing; the NMR data is collected by using high resolution (500MHz or more) spectrometers
- 3) Sequence-specific assignment: each peak in NMR spectrum corresponds to a particular atom (H, C, N etc) in the structure. These peaks are assigned to specific backbone or side chain residues in the sequence by using the chemical shift data and correlation methods e.g; correlation spectroscopy (COSY), total correlation spectroscopy (TOCSY), Heteronuclear single quantum coherence (HSQC) etc.
- 4) Finally structure calculation, model building and refinement

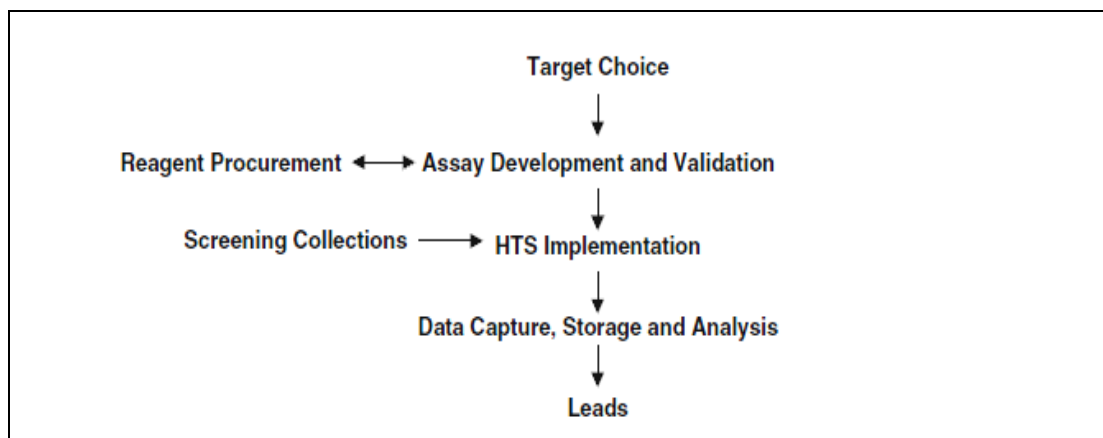
If the interest is only in the biophysical properties of the protein then only the chemical shift values for the backbone atoms of a protein need to be determined, while for NMR image of the entire protein, every atom in the protein needs to be assigned.

NMR spectroscopy has been used in numerous studies to assess the conformational dynamics and thermodynamics of protein-ligand binding [103]. In quest of the biological role of HI0719, over 100 naturally occurring small molecules or structural analogues were tested for ligand and some of the compounds were found to bind to the putative active site, which have previously been proposed to be involved in isoleucine biosynthetic pathway.  $^{15}\text{N}$ -HSQC spectra were recorded during the ligand screening by mixing test compounds with  $^{15}\text{N}$ -labelled HI0719 and comparing these spectra with the spectrum obtained from protein alone with the same buffer conditions. The extent of binding (in terms of  $K_d$  value) of particular ligand was determined by monitoring the change in chemical shifts of HSQC cross-peaks as a function of the ligand

concentration {104}. In another study NMR has been used to study the functional analysis of ligand-protein interactions, where ERB and its trans activation domain (TAD) and ligand binding domain (LBD) were examined for conformational changes in the absence and presence of 17 $\beta$ -estradiol. The  $^1\text{H}$ -NMR spectra showed that the LBD conformation was altered in the presence of 17  $\beta$ -estradiol, which in turn may help to develop therapeutic drugs to control the estradiol-mediated gene expression in hormone dependent diseases {105}.

### 1.3.3 High throughput screening (HTS)

In majority of the pharmaceutical and biotechnology companies, HTS is serving a fundamental role in drug-discovery process, in brief the HTS process is the journey from target choice eventually resulting in Lead discovery. HTS is a multi-disciplinary approach and requires strict adherence with respect to a number of factors. Certain steps and their organization in the course of HTS have vital importance in examining screening hits. To search for patterns which define lead series and to make the process successful the steps include biochemical assay methods, screening laboratories, screening robotics and for all these processes the technology specialists (chemist, biochemist, biologist, IT personnel) who are capable of maintaining, distributing, collecting and analyzing large datasets.



**Figure 1.16 Main steps involved in the HTS process from target choice to lead identification, image modified from {106}**

Through the combine efforts of all above mentioned and with the blend of various specialities, therapeutic targets can be passed through the lead discovery engine called HTS which can give rise to lead compounds (Figure 3.7). HTS is considered to be the heart of the drug discovery process. For the critical



design and implementation of HTS approach in a well organized manner it involves researchers from multiple disciplines of science. With the strict consideration of undergoing multi tasks, HTS starts with the choice of assay target and finally ends with the discovery of lead compounds. The outcome of results in the form of success and failure depends to a major extent on the decisions made during the whole process. In many instances the decisions have to be pragmatic and able to withstand the opposing forces {106}.

### 1.3.3.1 Enzyme Kinetics

With the help of industry and academia the development and optimization of fluorescent and/or absorbance based enzyme assays have performed a significant role in the screening process of hits. In HTS process in order to find initial actives, majority of the enzymes have been used as a drug target in the pharmaceutical industry. Generalized approaches are used for developing, validating, and troubleshooting assays which are applicable in both industrial and academic settings. Various enzyme classes including kinases, proteases, transferases, and phosphatases have been used to illustrate assay development approaches and solutions, which further lead to the identification of activators {107}. In various other examples the enzyme assays have been carried out in order to find suitable inhibitors, in the nano molar and pico molar range {108-111}. Apart from HTS, one of the main characteristics of enzymes is their absolute and relative specificity, in the former case the enzyme can only act on one substrate, while in the latter case the enzyme catalyzes some structurally related substrates and demonstrates relative specificity over other substrates. substrate specificity of the enzyme is crucial in enzyme assays and has major importance in identifying the capability of converting multi-substrates. Classic examples are of aldose reductase and amido hydrolases superfamily {14, 112}. In cases of specific enzymatic proteins pharmacophore searching process can be helpful in identifying sensible hits from the corresponding database. In order to test whether a hit is a potential substrate or not the hits could be routed to established enzyme assay.

### ***1.3.4 Isothermal titration calorimetry (ITC)***

ITC is the well characterized and widely accepted technique among the scientific community. It is used for the measuring of thermodynamic parameters like enthalpy, entropy and association constants. ITC is used to measure the heat change as a result of incremental addition of ligand to the target protein at a constant temperature. Any binding interactions between the ligand and target protein results in a temperature change. In ITC there is sequential addition of ligand to the protein at constant pressure. The energy change associated with the addition of ligand is the change in enthalpy. The added advantage of ITC over other techniques is that it measures the change in enthalpy directly [42].

The instrument consists of 2 cells. One is a reference cell, which has only buffer in it and the other is the sample cell containing a known concentration of target protein in a buffer solution. The temperature difference between these two cells is monitored and kept constant by the feedback heaters. Any change in temperature activates the feedback heaters to apply either increase or decrease in energy to the sample cell to match the reference cell. This ensures that the temperature difference between two cells is at minimum. The energy required to keep the two cells constant is called the baseline energy. Pre-selected volumes of the ligand are injected in both the cells. If the binding event is endothermic, the sample cell becomes cooler than the reference cell and energy is added by the feedback heater until the temperature becomes same as of reference cell and vice versa in exothermic binding event. The energy increase/decrease required to correct any imbalance is measured and recorded for numerous additions of known volumes of ligand until all the binding sites are occupied. Each of these deviations from the baseline are plotted as energy versus time resulting in a peak for each injection. The area of these peaks indicates the energy change from the baseline at each injection. This can be calculated by integration of the graph to give the heat evolved or absorbed during the binding event with respect to time. In order to get true binding energies same additions of ligand are carried out in the reference cell. The heat of dilution is calculated and subtracted. Finally a binding curve is produced through which various thermodynamic binding parameters, like binding constants ( $K_a$ ), reaction stoichiometry ( $n$ ), enthalpy ( $\Delta H$ ) and entropy ( $\Delta S$ ) can be extracted [97]. In this work ITC served as a highly sensitive technique,

specifically for the determination of strong binding events. Various studies have been carried out by using ITC as a biophysical tool, in some analysis the toxicity of certain compounds over specific proteins is investigated {113}, while in some cases the protonation changes in trypsin and thrombin upon ligand binding {114}. Recently in a novel ITC method the changes in thermodynamic parameters were studied during the folding of RNA coupled with magnesium uptake {115}.

### ***1.3.5 Differential scanning calorimetry (DSC)***

DSC is also used for the analysis of thermodynamic changes during the binding event of protein with the ligand. It measures the heat capacity for protein-ligand binding by incrementally changing the temperature of the system over a specified range. DSC measures the difference in heat energy uptake between a sample solution and appropriate reference solution (buffer) with increase in temperature. A typical experiment comprises one (or more) scans of the sample solution, together with control experiments using buffer alone to establish the instrument baseline. For quantitative analysis accurate sample concentrations are required, and parameters such as concentration, scan rate, pH, etc., may be changed. In DSC experiments the  $T_m$ , which appears as a transition peak on the DSC scan can be measured and is defined as the temperature at which one half of the protein is unfolded and the other half is folded. This can be related to the ligand-binding event as the ligand bound protein in solution will either be more or less thermally stable causing a change in the  $T_m$ . Typically the resultant increase in  $T_m$  demonstrates that the protein structure in the presence of ligand favors its folding and thus the stability of the protein {116}.

DSC has been an excellent tool in providing valuable information with respect to ligand binding events. It has been used to study the enthalpy changes coupled with the thermal unfolding and refolding of human serum retinol binding protein (sRBP), where it witnesses the formation of stable intermediates during the unfolding process {117}. DSC has also been used in clarifying metal-protein binding modalities {118} and for investigating the intrinsic protein structural energetics due to ligand-binding interactions {119}.

### ***1.3.6 Circular Dichroism (CD) spectroscopy***

Circular dichroism (CD) is a biophysical technique for studying protein structure in solution. CD spectroscopy is based on the principle to measure the difference in absorbance of right and left-handed circularly polarized light by chiral chromophores (typically aromatic amino acids) and to obtain structural information from the spectral features produced. The composition of secondary structure, such as  $\alpha$ -helices,  $\beta$ -sheets and turns can be estimated from the differential peptide bond absorbance. Various conformations give rise to characteristic CD spectra in the far UV region. The CD absorbance from aromatic residues in the near UV region provides information about the environment of these residues. Analysis of near UV CD spectra alone is not sufficient in provision of significant insight into tertiary structure but it does provide a distinctive fingerprint of the tertiary structure which can be further utilized to monitor structural changes {120}. CD spectroscopy is a versatile technique as it is helpful in providing structural information from different view points. It has been used to study the G-quadruplex folding patterns in STAT-3 which is the likely target for cancer therapy {121}. CD has been used to probe the effect of certain surfactants on structure and conformational stability of novel proteins coupled with pH changes {122}. To study the sequence effects on the over all structure of the protein as a result of mutagenesis {123} CD has provided reasonable insights. Although CD provides low resolution structural data compared with techniques such as X-ray crystallography and NMR, but its benefits such as time, sample concentration and cost favors it as a rapid screening technique for structural assessment of different proteins. The changes in CD spectra as a result of addition of ligand help in understanding the specific structural changes taking place in specific areas of the protein, which in turns helps to get some idea about ligand binding and its corresponding features. CD has a significant role to understand the undergoing structural and conformational changes at secondary structure level of the protein.

## 1.4 Aims and objectives of the project

The primary intention of the project is to optimize and apply a novel approach for using pharmacophores and pharmacophore-based screening in search of discovering the function and mechanism of action of target proteins. If successful the technique could be used as an aid for the discovery of protein function from structure. Apart from the interesting practical applications, the field of research is aimed to offer conceptual insight into the mechanism of molecular recognition between the proteins and the ligands.

Mainly the objectives include.

1. To address the use of pharmacophore searching for substrate prediction and hence functional assignment from structure.
2. To optimize the use of existing software to allow detailed pharmacophores to be generated and searched against a database of natural products. Therefore a number of known structures were chosen with known functions to validate the method and identify weaknesses and any problems with the method.
3. To take a known system with a number of ligands and potential ligands and assess the quickest and most effective methods to characterize binding.
4. It is proposed to use pharmacophore searching methods to interrogate the active sites cavities of a series of unknown protein structures with an aim to identify potential compounds that would help in exploring the potential function of the proteins.
5. Target proteins have a well defined active site cavity or cleft and contain sequence conserved amino acid residues. Enzymes with a predicted known functionality based on conserved catalytic residues will be considered where a suitable substrate has not been identified.
6. To use it in exploring vital interactions provided by the protein binding site.

## 2. MATERIALS AND METHODS

### 2.1 General reagents

All the analytical grade chemicals and biochemical's used were obtained from Sigma Aldrich® unless stated otherwise.

### 2.2 Bacterial strains

All the bacterial strains used during this PhD research work are listed in table 2.1

Escherichia coli	Relevant genotype	Notes
<i>E.coli</i> BL21 (DE3)	<i>F<sup>-</sup> ompT hsdS<sub>B</sub> (r<sub>B</sub><sup>-</sup> m<sub>B</sub><sup>-</sup>) gal dcm</i> (DE3)	General purpose expression host {124}
<i>E.coli</i> DH5α	<i>F- Φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17 (rK<sup>-</sup>, mK<sup>+</sup>) phoA supE44 λ<sup>-</sup> thi-1 gyrA96 relA1</i>	<i>recA1</i> ensures increased insert stability and prevents unwanted recombination. <i>endA1</i> improves the yield and quality of plasmid DNA prepared from minipreps {125}
<i>E.coli</i> BL21 Rossetta (DE3) Novagen®	<i>F ompT hsdS<sub>B</sub>(r<sub>B</sub><sup>-</sup> m<sub>B</sub><sup>-</sup>) gal dcm lacY1</i> (DE3) pRARE <sup>6</sup> (Cm <sup>R</sup> )	general expression host; <i>lac</i> permease mutation allows control of expression level, provides rare codon tRNAs {126}
<i>E.coli</i> Origami (DE3) Novagen®	<i>Δara-leu7697 ΔlacX74 ΔphoAPvull phoR araD139 ahpC GalE galk rpsL F{lac<sup>+</sup>(lacI<sup>q</sup>)pro} gor522::Tn10</i> (Tc <sup>R</sup> ) <i>trxB::kan</i> (DE3)	general expression host; two mutations in cytoplasmic disulfide reduction pathway enhance disulfide bond formation in <i>E. coli</i> cytoplasm {126}

**Table 2.1** Different type of E-coli cells used for over expressing protein and their relevant genotype

## 2.3 Plasmids

All the bacterial strains used during this PhD research work are listed in table 2.2

Plasmid	Source	Selection	Notes
pET-FG <sup>1</sup>	Prof. Dr.Gary Sawers, Martin-Luther University, Halle, Germany	Kanamycin	T7 expression plasmid with N-terminal 6xHis- tag, PDB code: 2UYN
pTBL2	Dr. Bernhard Lokhamp, University of Glasgow	Tetracycline	T7 expression plasmid with N-terminal 6xHis- tag, HutD, PDB code: 1YLL
pTBL2	Dr. Neil Patterson, University of Glasgow	Tetracycline	T7 expression plasmid with N-terminal 6xHis- tagged NFGase, HutG, (PAA)
HK100	Joint center for structural genomics (JCSG), The Scripps Research Institute, North Torrey Pines Road, La Jolla, California 92037	Kanamycin	T7 expression plasmid with Clone site: SpeedET, N-terminal 6xHis tag, PDBcode:2Q7S
pTB361	Dr. Kirsty Stewart, University of Glasgow	Tetracycline	<i>HlsdhL</i>
P1X-02	Chantal Abergel, CNRS, Aix- Marseille Université Marseille, France	Ampicillin	T7 expression plasmid with N-terminal 6xHis- tag, Candida albicans DHQase, PDBcode:3KIP
PAO1	Dr. Kirsty Stewart, University of Glasgow	Ampicillin	N-terminal 6x His tag, <i>Helicobacter pylori</i> Type II DHQase
pET15b	Dr. Kirsty Stewart, University of Glasgow	Ampicillin	<i>Streptomyces coelicolor</i> Type II DHQase
pT7.7	Dr.Lewis Evans, University of Glasgow	Ampicillin	<i>Mycobacterium tuberculosis</i> Type II DHQase

**Table 2.2 Details of plasmids, their source, selection and resultant protein**

## 2.4 Plasmid purification (mini prep)

The plasmid purification for plasmids was carried out as per protocol of QIAGEN® Miniprep Handbook, page 22, 2<sup>nd</sup> Edition, December 2006.

The protocol is designed for purification of up to 20 µg of high-copy plasmid DNA from 1-5 ml overnight cultures of *E. coli* in LB (Luria-Bertani) medium. All steps were carried out at room temperature.

1. Pelleted bacterial cells were resuspended in 250 µl Buffer P1 and transferred to a micro centrifuge tube.
2. 250 µl of Buffer P2 was added and mix thoroughly by inverting the tube 4-6 times.
3. 350 µl of Buffer N3 added and mix immediately and thoroughly by inverting the tube 4-6 times.
4. Centrifuged for 10 min at 13,000 rpm (~17,900 x g) in a table-top micro centrifuge.
5. The supernatants from step 4 were loaded on to the QIAprep spin column by decanting or pipetting.
6. Centrifuged for 30-60 s and the flow-through discarded.
7. The QIAprep spin column was washed by adding 0.5 ml Buffer PB and centrifuged for 30-60 s and the flow-through discarded.
8. QIAprep spin column was washed by adding 0.75 ml Buffer PE and centrifuged for 30-60 s.
9. The flow-through discarded, and centrifuged for an additional 1 min to remove residual wash buffer.
10. The QIAprep column was placed in a clean 1.5 ml micro centrifuge tube. 50 µl Buffer EB (10 mM Tris·Cl, pH 8.5) or water was added to the center of each QIAprep spin column to elute DNA. The column was let to stand for 1 min and then centrifuged for 1 min.



## 2.5 DNA quantification

The concentration of DNA plasmid was calculated by measuring the absorbance at 260 nm (A<sub>260nm</sub>) using the nanodrop spectrophotometer (Thermo Scientific). Concentration was calculated based upon the equation. DNA Concentration (ng/μL) = A<sub>260</sub>/0.02 {127}

## 2.6 pH measurements

All pH measurements were made with pH meter (pH 211, Hanna Instruments) unless stated otherwise. The pH meter was calibrated with pH buffers: 4, 7 & 9

## 2.7 Antibiotics:

Four different antibiotics were used namely Ampicillin, Tetracycline, Chloramphenicol and Kanamycin. Stock solutions of the antibiotics were prepared in given solvent (Table 2.3). If required the antibiotic solutions were sterilized by filtration through a 0.22 μm filter. The stock solutions were stored at -20°C.

Antibiotic	Concentration of stock solution (mg/mL)	Concentration used (μg/mL)	solvent	Filter sterilization
Ampicillin	100	50-100	Distilled water	Yes
Tetracycline	12.5	12.5	80% ethanol	No
Chloramphenicol	34.5	34.5	100% ethanol	No
Kanamycin	10	50	Distilled water	Yes

**Table 2.3 Different antibiotic stock solution and concentrations**

## 2.8 Culture media for bacterial growth

Different *E.coli* strains were used for growing bacterial culture in various culture media. Table 2.4 gives the composition of different growth media used.

Media	Composition per Liter	Instructions
LB (Luria-Bertani Broth) for cultures	10g Tryptone, 5g yeast extract, 10g NaCl	Sterilized by autoclaving at 15 psi for 30 minutes at 121 °C.
LB agar (for plates)	LB composition plus 15g micro agar	Sterilized by autoclaving at 15 psi for 30 minutes.
SOC Medium (for 100mL)	2g Tryptone, 0.5g yeast extract 1mL 1M NaCl, 0.25mL 1M KCl, 1mL Mg <sup>2+</sup> /glucose stock (1M MgCl <sub>2</sub> , 1M MgSO <sub>4</sub> , 2M glucose) filter sterilized	Dissolve all (except Mg/glucose stock) in 97 mL distilled water, autoclave, cool, add Mg <sup>2+</sup> /glucose stock, filter sterilize medium through 0.22 µm filter.
Minimal Media (M9)	<u>5X M9 (1L)</u>  Autoclave  64g Na <sub>2</sub> HPO <sub>4</sub> ·7 H <sub>2</sub> O OR 85.5g Na <sub>2</sub> HPO <sub>4</sub> ·12 H <sub>2</sub> O, 15g KH <sub>2</sub> PO <sub>4</sub> , 2.5g NaCl  <u>To make 500mL of M9</u>  100mL of 5X M9  <u>Plus Filter sterilize Media</u>  1mL 1M MgSO <sub>4</sub> , 1mL 50mM CaCl <sub>2</sub> , 1.5g Glucose, 0.5g NH <sub>4</sub> Cl, 400µL Thiamine (50mg/mL)	First Autoclave the 5X M9 media and then directly before use add the filter sterilized media to the cooled 5X M9 as per composition. Thiamine is optional but definitely improves growth rates. Use <sup>15</sup> N NH <sub>4</sub> Cl in case of labelled media preparation with the rest of recipe the same.
Auto Induction Media	<u>Autoclave</u>  Phosphate buffer pH:7.2 (6g Na <sub>2</sub> HPO <sub>4</sub> , 3g KH <sub>2</sub> PO <sub>4</sub> )  Tryptone 20g, Yeast Extract 5g, NaCl 5g  <u>Filter Sterilize media</u>  60% v/v Glycerol 10mL, 10% w/v Glucose 5mL, 8% w/v Lactose 25mL	Adjust the pH of phosphate buffer first and then add the media, autoclave to sterilize, leave it to cool, add the filter sterilize media just before use

**Table 2.4 Composition of different culture Media**

## 2.9 Preparation of ultra-competent E.coli (BL21 DE3) cells:

Competent cells were prepared based on method by Inoue {128} with some modifications. The protocol followed is given below:

50μL of BL21 (DE3) cells were taken from -80°C freezer and cultured in 125mL of (preincubated at 19°C) super optimal broth (SOB) in 500mL bowelled flask in duplicate. The cultures were incubated at 19°C in shaking incubator until OD ( $A_{600}$ ) reached 0.5 (normally it takes 24-36 hours). The flask was then placed on ice for 10min, and then the cells pelleted by spinning at 3000xg for 10 minutes at 4°C. The supernatant was discarded and the cell pellet was gently resuspended in 40mL of ice-cold transformation buffer (TB) with 0.7mL of dimethyl sulfoxide (DMSO, Stored at -20°C before use). The solution containing cells was aliquoted in 100 and 500μL portions in eppendorfs and frozen in liquid nitrogen. The cells were stored at -80°C. Composition of Media:

### 1. SOB solution:

0.5% yeast extract, 2% tryptone, 10mM NaCl, 2.5mM KCl, 10mM MgCl<sub>2</sub>  
10mM MgSO<sub>4</sub>

Final SOB solution was dissolved in double distilled water and autoclaved to sterilize.

### 2. TB solution

10mM PIPES, 15mM CaCl<sub>2</sub>, 250mM KCl

Final TB solution was dissolved in double distilled water and pH adjusted to 6.7 with KOH or HCl and then MnCl<sub>2</sub> added to 55mM. After adjusting the final volume the solution was sterilized by filtration with 0.45μm filter and then stored at 4°C.

The newly prepared competent cells were tested by transformation and the number of colonies seen gave an indication of competency of the cells.

## 2.10 Transformation protocol

The transformation protocol was modified from Novagen pET system manual {126}. Briefly the required competent cell tubes were taken from the  $-80^{\circ}\text{C}$  freezer and immediately placed on ice in such a manner that whole cell tube was immersed in ice except the cap. The cells were allowed to thaw on ice for 4-5 minutes and then visually examined to see if thawed. The tube gently finger-flicked 1-2 times to evenly resuspend the cells. 50 $\mu\text{L}$  aliquot of cells were pipetted out in pre-chilled eppendorf tubes on ice and then 2.0  $\mu\text{L}$  of the required plasmid DNA was added and stirred gently to mix and then left on ice to incubate for 5 minutes. The eppendorfs tubes were then heated for exactly 30 seconds in a  $42^{\circ}\text{C}$  water bath without shaking and then put back on ice for 2 minutes. 250  $\mu\text{L}$  of room temperature SOC was added to the eppendorf tubes whilst on ice. The tubes were then incubated at  $37^{\circ}\text{C}$  while shaking at 250 rpm for 60 minutes in shaking incubator. The cells in eppendorf were poured on to pre-warmed agar plates containing the appropriate antibiotic and spread uniformly on the plates with the help of a sterilized glass rod. The plates were left on the bench for several minutes to allow maximum absorption of the liquid and then inverted and incubated overnight in a  $37^{\circ}\text{C}$  incubator. Same procedure was carried out for a control plate except the addition of plasmid DNA. The incubated plates were visually inspected for bacterial colonies. Whole procedure was carried out in sterilized bio hazard area with an open flame to prevent risk of bacterial cross contamination.

## 2.11 Storage of bacterial Strains

On temporary basis the selective antibiotic resistant agar plates with the bacterial strain were sealed in self sealing (nesco film) tape and stored in a cold room at  $4^{\circ}\text{C}$  for up to 4 weeks. For permanent storage the bacterial strains from overnight cultures were mixed with 100% glycerol to a final concentration of 80% glycerol/culture (v/v) and stored in  $-80^{\circ}\text{C}$  freezer.

## **2.12 Protein over-expression**

### **2.12.1 *Overnights and test of expression***

An individual bacterial colony was picked from an agar plate by using sterile inoculation loop and mixed in 10 mL of LB medium (with required antibiotic) in 50mL culture tubes and incubated over night in a 37°C shaking incubator. Next day these overnights were used for test of expression by seeding them to 10mL of LB medium in 1:10 dilution at above mentioned conditions. OD<sub>600</sub> of the cultures was monitored hourly and when reached between 0.6-0.8, the cultures were induced by addition of 1mM Isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG). 1mL samples were collected after each hour up to 4 hours duration. The samples were later on prepared for sodium dodecyl sulphate poly acrylamide gel electrophoresis (SDS PAGE) to check for protein over expression.

### **2.12.2 *Large scale growth of bacterial cultures***

An individual bacterial colony was picked from an agar plate by using sterile inoculation loop and mixed in 10 mL of LB medium (with required antibiotic) in 50mL culture tubes and incubated over night in a 37°C shaking incubator. Next day these overnights were used for large scale growth by seeding them to 250mL of LB medium in 2L flasks. OD was monitored hourly and when reached between 0.6-0.8, cultures were induced by addition of 1mM IPTG. These cultures were incubated overnight at 13°C. Same Protocol was followed when growing cultures in M9 media. In case of auto induction media cultures were grown by inoculating a single colony in 250mL of Auto induction media in 2L flask and incubated overnight in a 37°C shaking incubator.

### **2.12.3 *Cells harvesting***

The cells were harvested by centrifuging them for 25 minutes at a speed of 3000xg in a sigma® 4K15 centrifuge at 4°C. The bacterial cell pellet was resuspended in 15mL of binding buffer and stored at -20 °C.

### **2.12.4      *Cell lysis by ultrasonication***

Large scale cell lysis was carried out by using a sonicator. The frozen pellet was thawed at room temperature on a rotating stage. After thawing the sample a few mg of DNase were added and the tube was placed in slushy ice in a beaker to maintain the temperature at 4°C. The probe of the sonicator was fully immersed in the sample and the sonicator set for 30 cycles (each cycle = 30 seconds off, 30 seconds on). The cell lysate was then centrifuged for 25 minutes at a speed of 20000 rpm in a 3K30 sigma® centrifuge at 4°C. After centrifugation the supernatant was retained and stored on ice before subsequent purification steps.

## **2.13 Protein purification**

### **2.13.1      *Ni-NTA affinity chromatography or IMAC***

Ni-NTA affinity chromatography or immobilized metal affinity chromatography (IMAC) is a well established widely used one step purification technique for genetically engineered histidine tagged proteins. In this technique the stationary phase is mostly organic resin covalently bonded with a chelating ligand e.g; Nitrilo tri acetic acid (NTA), Imino di acetic acid (IDA) or tris-carboxymethyl ethylene diamine (TED). The ligands have the ability to form coordination complexes with Metal ions e.g; Cu(II), Zn(II), Ni(II), Fe(II). The chelating ligands have electron-pair donor atoms like N, S, O and the metal ions are electron-pair acceptor atoms. When the metal ion solution ( $\text{Ni}_2\text{SO}_4$ ) is passed through the stationary phase the formation of metal-ligand complex takes place, which serves as affinity ligand for the His-tagged proteins. The metal ligand complex can be bidentate, tridentate, etc., as per capacity of ligand to form coordination bonds. The rest of the coordination bonds are formed with water molecules in the beginning which are subsequently replaced by histidine residues of the recombinant protein {129}. The imidazole Nitrogen of histidine binds to the  $\text{Ni}^{2+}$ . To ensure the binding of target protein to the column the procedure has to be carried out at a neutral or slightly basic pH at which the imidazole Nitrogen of the Histidine residues is in nonprotonated form {130}. The cell lysate supernatant obtained after subsequent sonication, centrifugation and filtration was loaded on the Ni-NTA column pre-equilibrated with the binding buffer (Buffer A). The non his-tagged proteins are passed through the column

without binding while non-specific proteins which have few surface exposed histidine residues, weakly bind to the column and are removed by washing buffer (Buffer B) which contains slightly higher amount of imidazole. The target protein is eluted by changing the buffer conditions, mainly by increasing the imidazole concentration up to 300mM, so that it competes with the imidazole of the histidine tag and elutes it from the column. Details of the buffer

Binding buffer, A = 5mM Imidazole, 50mM Tris-HCL, 300mM NaCl pH: 7.6

Wash buffer, B = 20mM Imidazole, 50mM Tris-HCL, 300mM NaCl pH: 7.6

Elution buffer, C = 300mM Imidazole, 50mM Tris-HCL, 300mM NaCl pH: 7.6

### **2.13.2 Gel filtration chromatography**

Size exclusion chromatography was carried out by using a pre-packed Superdex 75 10/300 GL column (GE healthcare, #17-5174-01). The column was connected to the AKTA® chromatography system inside a cold room (AKTA Laboratory-scale Chromatography Systems provided by GE life sciences is for liquid chromatography systems which is used for protein purification at research laboratory scale. Systems provide more control than manual purification because of the ability to automatically control the flow rate and monitor the progress of the purification as well as to make controlled gradients and automatically collect fractions). After equilibration with 30 mL of running buffer at a flow rate of 0.5 mL/min, a 500 µL protein sample was injected into the sample loop. Size exclusion chromatography was then carried out at 0.5 mL/min running buffer for 1.5 column volumes (35.34 mL) and was monitored by the UV absorbance at 280 nm. Fractions of 0.5 mL eluant were collected and those which showed absorbance at 280 nm were analysed by SDS-PAGE. All the buffers used for gel filtration chromatography were first filtered and then degassed. The typical composition of the buffer for Gel filtration chromatography was 20mM Na Phosphate, 150mM NaCl, 0.01% Na-azide pH: 7.5 unless stated otherwise.

## **2.14 Dehydroquinases purification protocol**

The protein samples were purified by using the subsequent purification steps involving, ion exchange chromatography, hydrophobic interaction

chromatography (HIC), and finally Gel filtration chromatography. The details for ion exchange and Hydrophobic interaction chromatography, which were set through the AKTA® system, are given below.

### **2.14.1      *Ion exchange chromatography***

After centrifugation the supernatant from sonicated samples was loaded on to a Q-sepharose (anion exchanger) column which had previously been equilibrated with Buffer A (50mM Tris HCl, 1mM DTT pH: 7.8). A linear gradient of 0-600mM NaCl of buffer B (50mM Tris HCl, 0.6M NaCl, 1mM DTT pH: 7.8) was applied to elute the protein. 5 mL fractions were collected and analyzed for enzyme activity. Fractions containing maximum activity were pooled. For the precipitation of proteins (*HlsdhL*), 3M ammonium sulfate solution was added to the pooled fractions to a final concentration of 1.0M. The samples were spun at 20000xg for the removal of precipitated protein. Finally the solution was concentrated to 10mL using an AMICON® protein concentrator with a 5000kDa molecular weight cut off (MWCO).

### **2.14.2      *Hydrophobic interaction chromatography (HIC)***

The samples obtained through ion exchange chromatography were loaded on the phenyl-sepharose column with a flow rate of 1mL/min. The column was pre-equilibrated with buffer B (1.4M (NH<sub>4</sub>)<sub>2</sub> SO<sub>4</sub>, 50mM Phosphate, 1mM DTT pH: 6.8). Initially the column was washed with 10 column volumes of buffer B and then a linear gradient from 1.4-0.0 M (NH<sub>4</sub>)<sub>2</sub> SO<sub>4</sub> was applied along with buffer A (50mM Phosphate, 1mM DTT pH: 6.8). 5mL fractions were collected concentrated and then analyzed for enzyme activity. Fractions containing maximum activity were pooled and concentrated down at 20000xg and then loaded on to gel filtration superdex 200 column as a final purification step.

### **2.14.3      *Gel filtration chromatography on superdex 200 column***

The pooled fractions obtained through HIC were loaded on superdex 200 (gel filtration) column equilibrated in buffer C (50mM Tris HCl, 200mM NaCl, 1mM DTT pH: 7.8) at a flow rate of 2mL/min. The Protein was eluted using buffer C. 2mL fractions were collected. Collected fractions were pooled and subjected to enzyme assays for checking activity.



## **2.15 Protein characterization**

### **2.15.1 SDS-PAGE**

SDS-PAGE is a routine method to detect protein expression and purity. The NuPAGE<sup>®</sup> system (Invitrogen), consisting of 12% Bis-Tris Novex Mini-gels and an Xcell SureLock<sup>™</sup> Mini-Cell was used for all SDS-PAGE. The system was assembled according to the manufacturer's instructions.

### **2.15.2 Sample preparation for SDS Gel Electrophoresis**

In case of test of expression, 1mL samples of the induced and uninduced bacterial cells were centrifuged. The supernatant and the pellet were retained. The pellet was resuspended in 50 $\mu$ L (equivalent to of 1/20<sup>th</sup> of the original sample volume) of detergent mix "bug buster" using vortex. The samples were left for 20 minutes at room temperature; distilled water was added dependent on the OD of the original sample to give consistent loading of protein on the gel. From each sample, 30 $\mu$ L of sample was added to new eppendorfs containing 10  $\mu$ L of loading buffer (which contains  $\beta$ -mercaptoethanol, the detergent SDS, and a marker dye to follow the progress of the electrophoresis) and heated for 5 minutes at 85°C on a heat block to denature the proteins. 10  $\mu$ L of prepared protein sample was loaded onto wells in the precast gels and molecular weight markers were added in certain wells for reference purposes.

### **2.15.3 Staining and destaining procedure**

Analysis of the electrophoretic profile was performed by staining the gel with coomassie stain (1 g coomassie brilliant blue R-250, 50 ml glacial acetic acid, 500 ml MeOH, 450 ml H<sub>2</sub>O). The gel was incubated in 100mL Coomassie stain for 20 minutes. The background stain was removed by immersion of the gel in 200mL of destaining solution (10%glacial acetic acid, 10% methanol) for overnight. Images of the gel were taken by using the Kodak imager<sup>®</sup>.

### **2.15.4      *Measurement of protein concentration via UV absorbance***

The A<sub>280</sub> absorbance of protein solutions were measured by either using a Nano Drop 1000 (Thermo Scientific) or the JASCO V-550 spectrophotometer. It was noticed that the JASCO V-550 spectrophotometer gave more accurate results than Nano Drop 1000 therefore former was used when more accuracy was needed. Extinction co-efficients were calculated from amino acid sequence {131}. Routinely this was carried out by using the ProtParam program (<http://web.expasy.org/protparam/>). The Beer-lambert law was also used to calculate the protein concentration and extinction coefficient values from A<sub>280</sub> absorbance.

### **2.15.5      *Dialysis***

It is necessary to completely exchange the buffers a protein is in for a number of applications e.g; enzyme assay, crystallization. The protein solution was added to the dialysis tubing and then sealed at both ends with plastic clips. The sealed samples were then suspended in a large volume (1-2 Liters) of suitable buffer for overnight on a magnetic plate with constant stirring in cold room at 4 °C.

### **2.15.6      *PD-10 desalting column (Buffer exchange)***

Another method for buffer exchange was by using a PD-10 column. Size exclusion chromatography (gel filtration) is a gentle technique for the purification of protein sample, based on the size of pores of column. The large molecules are eluted first and the small one's at the end. PD-10 desalting column was equilibrated with a 5 column volumes of a suitable buffer. 2.5 mL of protein sample was added and the flow through was discarded. Finally the elution with 3.5 ml of elute buffer.

### **2.15.7      *Lyophilization (Freeze drying)***

Protein samples were dialyzed and then concentrated in a suitable buffer to a volume of 1-2mls by using centrifugal filter unit (centricon®) in a centrifuge (sigma 3K30). The concentrated protein was placed in a round bottom flask and flash frozen in liquid nitrogen. The samples were swirled in the flask to

maximize the surface area of the frozen solid and thereby speedup freeze drying. The frozen sample was attached to the freeze dryer and subjected to vacuum drying overnight at  $-110^{\circ}\text{C}$ . Upon completion of freeze drying, solid protein was scrapped from the flask and stored at room temperature.

## 2.16 Protein crystallization

The use of crystallization technique for the determination of structure of protein is now widely and frequently used, since the protein structure of hemoglobin and myoglobin were determined by Perutz and Kendrew {132} {133}. The number of protein structure determined by crystallography is increasing day by day as of November 2012 there are some 86,334 structures held in the Protein data Bank (PDB) {19}. Protein crystals were grown using the sitting drop vapour diffusion method in a Crystchem<sup>TM</sup> 24 well plate (Hampton Research). The set up for sitting drop vapour diffusion method was followed as:

Crystchem Plates<sup>TM</sup> (Hampton Research Ltd) were used to set up crystallizations with up to 24 different conditions per plate. The well was prepared first and usually contained 1ml of a buffered precipitant solution such as polyethylene glycol or ammonium sulfate or a mixture of PEG and salt. Sometimes additives were also included such as detergents or metal ions which may enhance the crystallization. 1 $\mu\text{L}$  of the concentrated protein sample was pipetted onto a concave platform, followed by 1 $\mu\text{L}$  of the well solution. The tray was then sealed over by using a crystal clear tape. This was then left undisturbed for at least 24 hours to equilibrate. At the start of the experiment, the precipitant concentration in the drop is half to that of the well. Equilibration then takes place via the vapor phase. Given the relatively large volume of the well, its concentration effectively remains the same. The drop in the concave platform however loses water vapors to the well until the precipitant concentration equals to that of the well. A wide range of crystallization conditions were used by varying the precipitant, buffer and additives concentrations. The pH levels types of precipitant and additives were also changed to get bigger crystals. The crystallization screen was left at  $20^{\circ}\text{C}$  for an indefinite amount of time. Screening for crystal growth was intermittently done by using an Olympus (model: S240) microscope.

### **2.16.1      *Streak seeding***

The technique for assisting the crystallization of proteins was carried out by dipping cat hair in to the drop which has crystalline precipitates. The cat hair was then streaked onto other clear drops and left for growth of individual crystals.

## **2.17 Circular dichroism (CD) spectroscopy**

Circular Dichroism (CD) experiments were performed by Dr Sharon M Kelly (Protein Characterization Facility, University of Glasgow) using a JASCO J-810 spectropolarimeter. Near UV CD spectra were recorded in quartz cuvettes of 0.2cm or 0.5cm path lengths using the following parameters: scan rate 10nm/min; response 2 sec; bandwidth 1.0nm; no. of scans 3. Far UV CD spectra were recorded in quartz cuvettes of 0.02cm path length using the following parameters: scan rate 50nm/min; response 0.5 sec; bandwidth 1.0nm; no. of scans 6.

## **2.18 Isothermal titration calorimetry (ITC) and differential Scanning calorimetry (DSC)**

Isothermal Calorimetry (ITC) and Differential Scanning Calorimetry (DSC) experiments were performed by Margaret Nutley (Department of Chemistry, University of Glasgow) using Microcal VP-ITC and VP-DSC instruments.

## **2.19 Enzyme Assays**

All enzyme assays were carried out using Jasco V-550 dual beam spectrophotometer. Typically a 1mL quartz cell with 1cm path length or if absorbance was too high a 0.5cm or 0.2cm path length cells were used. Measurements were made in triplicate at a wavelength of 234nm for DHQases and at 210nm for NFGases. The data obtained through enzyme assays was processed in Microsoft Excel or Origin software to calculate the  $V_{\max}$ ,  $K_m$  and  $K_{cat}$  values.

### 2.19.1 *Standard 3-dehydroquinase assay*

The Standard assay to measure 3-dehydroquinase activity follows the formation of 3-dehydroshikimate. Dehydroquininate is added to the protein sample, and the rate of conversion to 3-dehydroshikimate is monitored by the increase in  $A_{234}$  absorbance. Enzyme assays for all the Dehydroquinases were carried out in 50mM Tris buffer, pH: 7.6 at room temperature. The final concentration of protein in the cell was 0.3 $\mu$ M. The concentration of substrate (Dehydroquininate) in the cell was used in the range of 9.0–350  $\mu$ M. The increase in absorbance at 234nm was measured for 120 seconds. The  $K_m$ ,  $V_{max}$  and  $K_{cat}$  values were calculated by using the Michaelis–Menten kinetics (Eq: 1) and Lineweaver–Burk plot (Eq: 2).

$$v = \frac{V_{max}[S]}{K_m + [S]} \quad \text{Eq:1}$$

$$\frac{1}{v} = \left[ \frac{K_m(1)}{V_{max}[S]} + \frac{1}{V_{max}} \right] \quad \text{Eq:2}$$

### 2.19.2 *Standard NFGase assay*

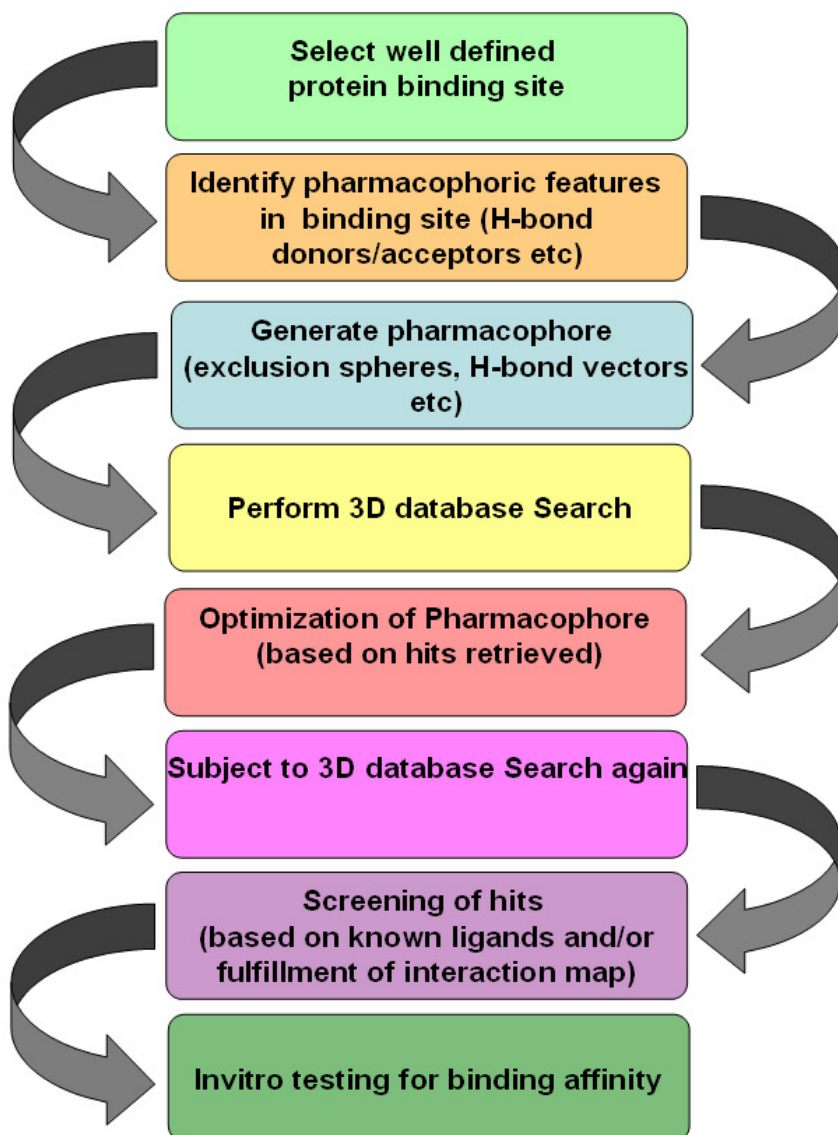
The Standard assay to measure NFGase (N-formyl glutamate amido hydrolase) activity follows the hydrolysis of N-formyl L-glutamate to formate and glutamate. The assay follows the hydrolysis of the amide bond and is monitored by the decrease in  $A_{210}$  absorbance. The enzyme assays for NFGase were carried out in PBS (Phosphate buffer saline; 137mM NaCl, 2.7mM KCl, 10.0mM  $\text{Na}_2\text{HPO}_4 \cdot 2\text{H}_2\text{O}$ , 2.0mM  $\text{KH}_2\text{PO}_4$ ) buffer, pH: 7.4 at room temperature. The concentration of substrate (N-formyl L-glutamate) in the cell was used in the range of 0.05–1.5 mM. The final concentration of protein in the cell was 2.0 $\mu$ M. The decrease in absorbance at 210nm was measured for 120 seconds. The  $K_m$ ,  $V_{max}$ ,  $K_{cat}$  and  $K_i$  values were calculated by using the Michaelis–Menten kinetics and Lineweaver–Burk plot.

### **3. Pharmacophore methodology**

#### **3.1 Overview of pharmacophore methodology**

The generation of a pharmacophore model from a protein structure involves the following steps:

1. Identification of the protein active site and the selection of a sphere of atom coordinates that describe it.
2. Visual identification of structural and chemical features within the active/binding site of the protein.
3. Visual recognition of potential H-bond donor and acceptor interactions from the protein perspective.
4. Selection of protein atoms within the active site to act as exclusion spheres limiting the conformational space available to any ligand.
5. Search of the model pharmacophore against a database of compounds.
6. Assessment of both of the number and quality of the hits obtained from pharmacophore searching, optimization of the pharmacophore and iteration of the procedure.
7. Visual check of the hits obtained with naked eye, regarding their proper orientation in the active site, ruling out the possibility of steric clash and satisfying the constraints, primarily in terms of H-bond donor and acceptor interactions.



**Figure 3.1** Flow chart describing the generalized scheme of events in the generation and optimization of a pharmacophore

The speed of the process, both in generation of the model as well as the prompt search results from minutes to a few hours is an added advantage of pharmacophore searching over other methods such as docking. Among other advantages is the flexibility in the generation of the pharmacophore model. The ability to test hypotheses and further improvement of the model based on initial hits obtained from preliminary searches. An important feature is the freedom to vary the bond lengths and bond angles of the H-bond vector interactions within the allowed tolerances known for the interaction. The following computer

programs, databases and methodologies were used during the course of pharmacophore searching process for various proteins.

## 3.2 Catalyst®

Catalyst (Accelrys) version 4.9 was used to check for errors in screening the hypotheses against the required 3D conformational database by using the “fast flexible search” option available in the catalyst stockroom database tools window. Either too many hits or more often no hits would indicate problems with the generated hypothesis which could be addressed. Default parameters were used throughout unless otherwise stated.

Once the pharmacophore had been checked for correct format and for other errors, the catSearch option was run as a shell script in the form of following command: **sh multi2.sh**

Two types of output in the form of .sd file format were obtained from searches using the following catSearch commands

1. AllHitConfs caused all the conformers that can be mapped to the query:

```
catSearch $1 -query $i -maxhits 300 -allHitConfs -align -sd $k
```

2. BestFlexible which performed best flexible search to the query:

```
catSearch $1 -query $i -maxhits 300 -bestFlexible -align -sd $k
```

The multiple conformations (AllHitConfs) option helped in finding the most favored orientation of the key functional groups of the ligand towards the binding site. However, the Best Flexible Hit was used most often to limit the number of Hits to a manageable number for visual inspection.

In later more sophisticated searches; Catalyst was used to optimize certain parameters like type of bond between query atoms, charge on atoms and multiple atom types for individual query atoms of the pharmacophore model. Searches were carried out by using catalyst for multi-conformational database of compounds. The resultant pharmacophore readout was a 3D map with respect to



the protein structure, representing the optimum number of geometric constraints. The model defined the essence of the target protein binding site and resulted in sensible hits.

### 3.3 Accelrys Discovery studio visualizer<sup>®</sup> (DSV)

As the optimization of pharmacophoric features with respect to the protein binding site and ligand attributes, requires a detailed understanding of the intermolecular interactions, DSV is a powerful tool for in detail visual inspection of such interactions. DSV was used in the generation of the later more sophisticated pharmacophores in Vector and Query atom methods. This software permitted the generation of the hypotheses without the use of Cerius2 and packages within this interface which resulted in significant savings in time and effort.

### 3.4 Generation of Databases

Following databases were generated and used throughout the project unless otherwise stated.

#### ***3.4.1 Generation of Aldehyde and ketone compounds database***

The database was created by using the Accelrys Catalyst Tutorials (Release 4.7), the generation of the database involved the following steps:

1. Generation of aldehyde and ketone functional group queries by using the hypothesis feature of the catalyst<sup>®</sup> software.
2. Merging of the hypothesis by using the Exclude/OR Edit option of the catalyst software (Catalyst<sup>®</sup> Tutorials, page 327)
3. Searching the hypothesis against available database (naturalism.bdb) by using the Best flexible search databases/spreadsheets option in Catalyst: stock room
4. Exporting the resulting output spreadsheet in the form of MDL structure data (SD format) to the home directory

5. Conversion of spreadsheet (.sd file format) to database file (.bdb file format) by using these commands (Catalyst Tutorials, page 310 & 413)

(i) `catDB CONFIG name of SD file .bdb` (this commands builds the database from SD file, and configures the new database by using the default database feature dictionary)

(ii) `catDB SD defaultdatabase.sd name of SD file.bdb` (this command creates the database from the SD file)

### ***3.4.2 Generation of dipeptide and tripeptide database***

The dipeptide and tripeptide databases were generated by initially using a shell script to generate a string of peptide descriptions using three letter codes for amino acids. These were interpreted and converted using molconvert from the JChem suite of programs, resulting into .sd format. The .sd files for di-, and tripeptides were combined into a single multi compound .sd file which was converted to database format (.bdb) as described earlier in section 3.4.1 (5). This resulted in 400 dipeptides and 8000 tripeptides resulting in the 8400 peptides in the peptide database. Other databases used were naturalism.bdb (5,492 compounds) a database of smaller compounds (Mr<500) from ChEBI {134} and the entire ChEBI database (16,799).

## **3.5 Pharmacophore generating methods**

A total of three distinct methods were used in the successive optimization of pharmacophore generation, dependent on the available software at the time.

1. Cerius2, notepad and WebLab viewer pro method
2. DSV, Vector method
3. DSV, Query atom method

The methods are described in detail in sections below.

### 3.5.1 Cerius2, Notepad and WebLab viewer pro method

This method was used in the preliminary stages of the project as a modification of the standard pharmacophore searching. The method was preferred for proteins whose structure with a ligand had been determined. During the course of optimization it was found very difficult to select potential interactions from the interaction map of the protein binding site. The method comprised of the following major steps:

- 1) Visualizing and optimizing active site of the protein to be probed with respect to the x y z coordinates by using Cerius2 program (Figure 3.2) and generation of hydrogen atoms. These required modifying the bond order for cofactors such as NADP+ or ADP.

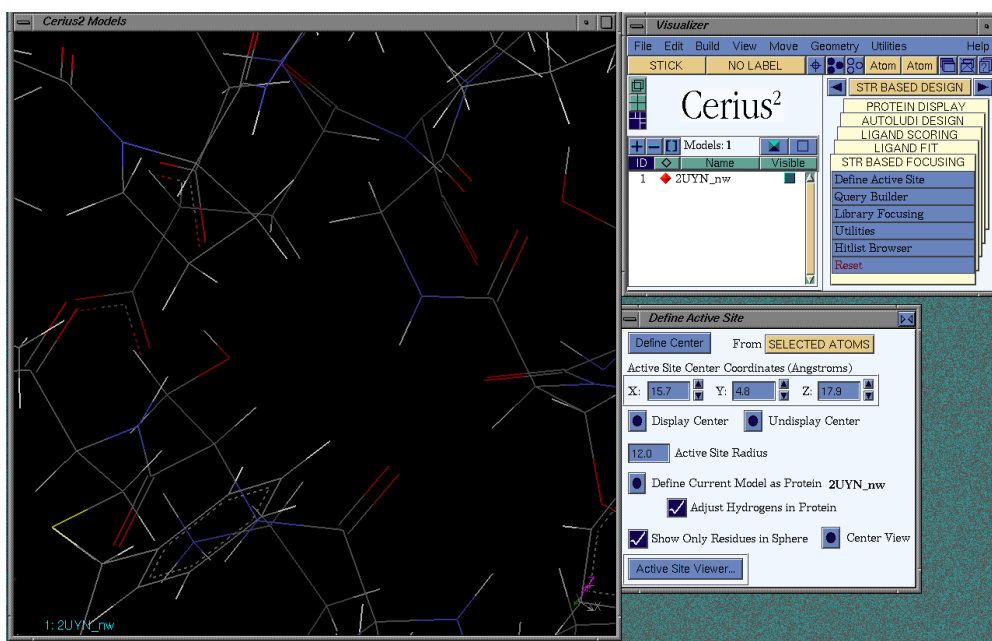
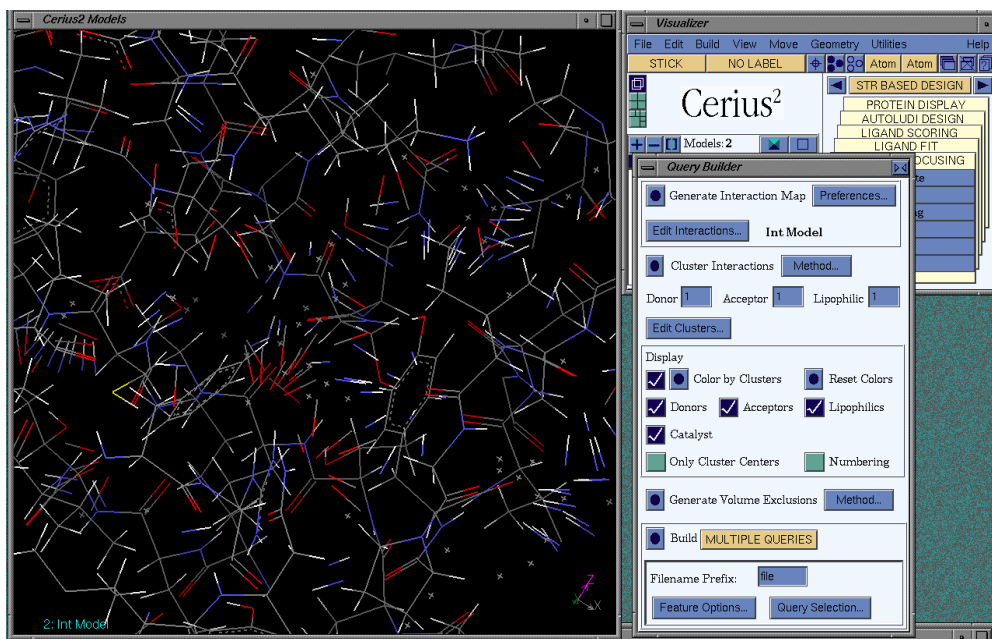


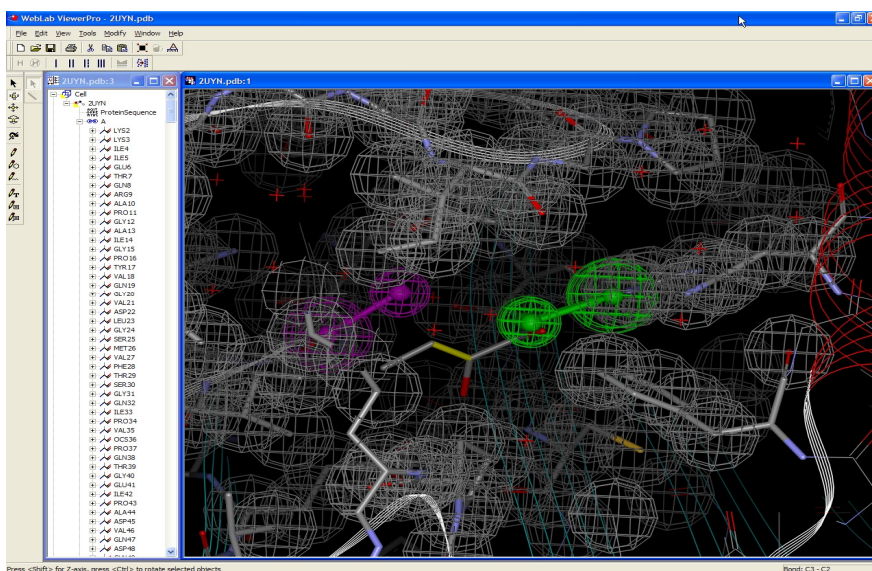
Figure 3.2 Visualization and optimization of protein active site by using Cerius2

- 2) Generation of interaction map and selection of potential interactions (donor, acceptor, lipophilic) manually as clustering of interactions did not produce reliable vectors. Further generation of exclusion spheres from protein atoms within the protein binding site normally of a radius of  $\sim 10\text{\AA}$  (Figure 3.3)



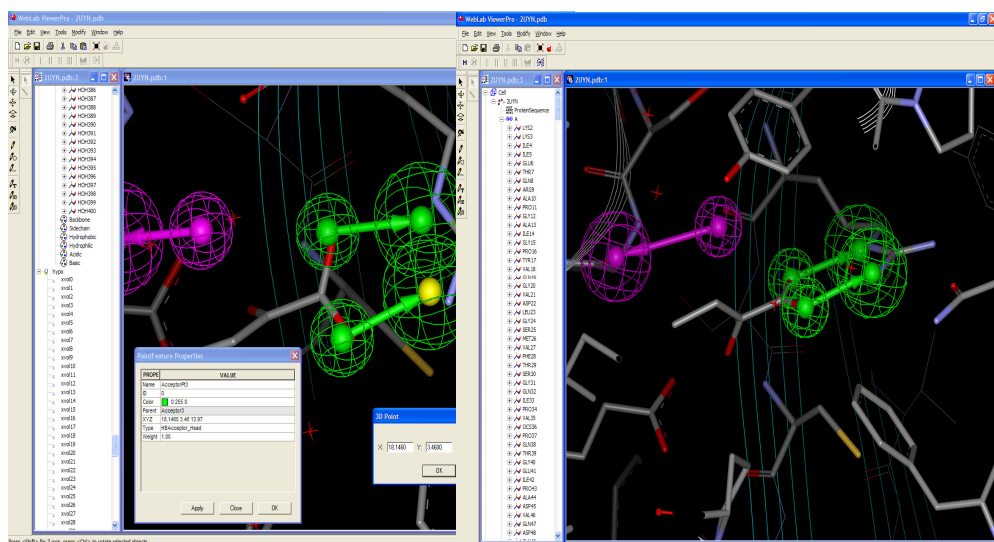
**Figure 3.3** Generation and optimization of potential interactions in the binding site (In Cerius 2 the small grey red arrows specify the possible H-bond interactions of the amino acids towards the potential ligand, important arrows responsible for the interactions were selected)

- 3) Visualizing the generated pharmacophore in Web lab viewer pro as this was more flexible and could be done easily together with the pharmacophore (Figure 3.4). In a suitable radius around the binding site each atom of the protein was provided with an exclusion sphere (grey spheres) to avoid the steric clash of the hits with the amino acids surrounding the binding site.



**Figure 3.4** view of generated pharmacophore in WebLab viewer pro (big grey spheres represent the exclusion volume, while the purple and green vectors represent the donor and acceptor interactions in the binding site)

- 4) Optimization of H-bond vectors direction by changing the x, y, z coordinate position in relation to the active site residues in web lab viewer pro (Figure 3.5) (exclusion spheres are hidden for clarity). The change in position of H-bond vectors was carried out in a following way: First note down the x, y, z co-ordinates for the desired position of the H-bond donor, then open the .chm file in word pad and change the x, y, z co-ordinates of the head and tail of the targeted H-bond vector (donor/acceptor) to the desired x, y, z co-ordinates and save the file.



**Figure 3.5 optimization of position of H-bond vectors in the binding site (The uncertainty spheres around the H-bond vectors define the tolerance limits for the hits which can be increased or decreased as per requirement)**

- 5) Using Catalyst and CatDB searches to search the pharmacophore against the multi-conformational compounds database.
- 6) Visualizing the results in WebLab viewer Pro

The procedure typically could be iterated 2-3 times as certain combinations of certain interactions will result in no compounds while others will result in 100s. Further the active site can be seen in web lab viewer pro along with the obtained hits. Manually certain hits can be rejected or accepted based on their interactions in the binding site. The addition of new vectors required could be carried out either by going back to Cerius2 or editing the .chm files by hand. This method did not permit the quick and easy evaluation and optimization of

the Pharmacophore. Equally, the difficulty to select chemical types such as a carboxylate was frustrating as this led to many false positives.

### **3.5.2 DSV, Vector method**

The major advantage of this method is the exclusion of the slow Cerius2 step and the generation of the pharmacophore within DSV. The step-wise detail of the methodology used for the generation of pharmacophore with a ligand in the active site is given below. Equally, water molecules could be used as sources of hydrogen bond donors or acceptors in the absence of ligand. During the course of optimization care was taken in satisfying the overall matching of the desired hydrogen bond in protein ligand interactions. Complementarity between the ligand and protein interface in terms of spatial occupancy and absence of any unfavorable interactions conferring steric conflict between the ligand and the binding pocket of the protein were also dealt with caution .

#### **3.5.2.1 Preparing the PDB file**

- 1) Generate hydrogens for the pdb file both for docking and for pharmacophore searching. This is best done using the What If Web Interface. <http://swift.cmbi.ru.nl/servers/html/index.html>  
Navigate using side menu to **Build/check/repair model** (Figure 3.6)

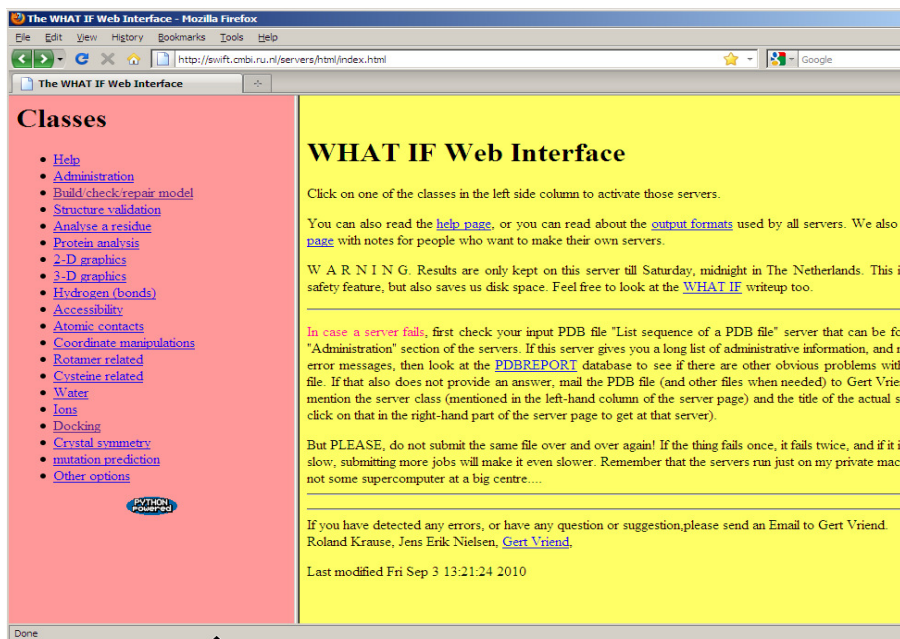


Figure 3.6 Image of web interface for what if program

- 1) Select from options in main window [Prepare PDB file for docking programs](#)
- 2) Prepare PDB file window allows to upload a PDB (Figure 3.7), obtained from the PDB database

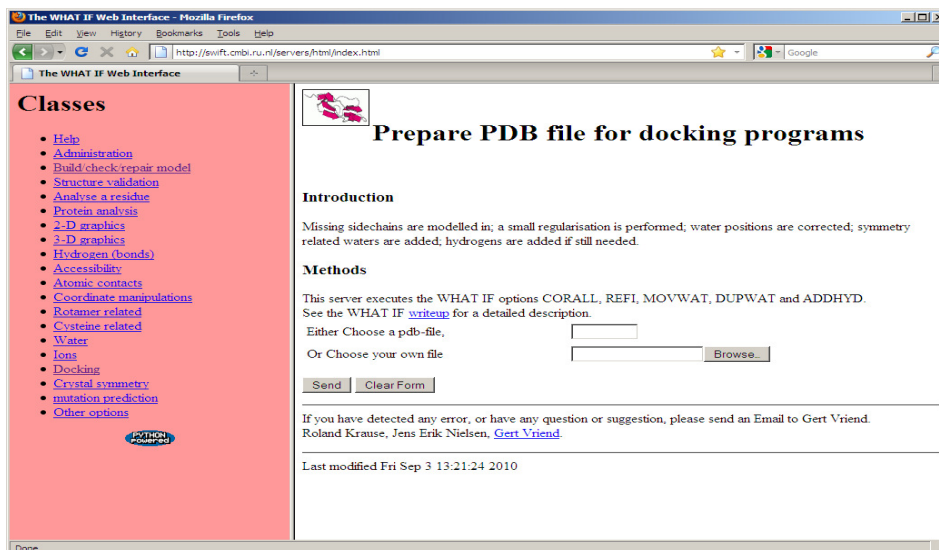


Figure 3.7 Image for what if weblink for uploading PDB file

- 3) Once uploaded a PDB the program will take a few minutes to get to the following page (Figure 3.8).



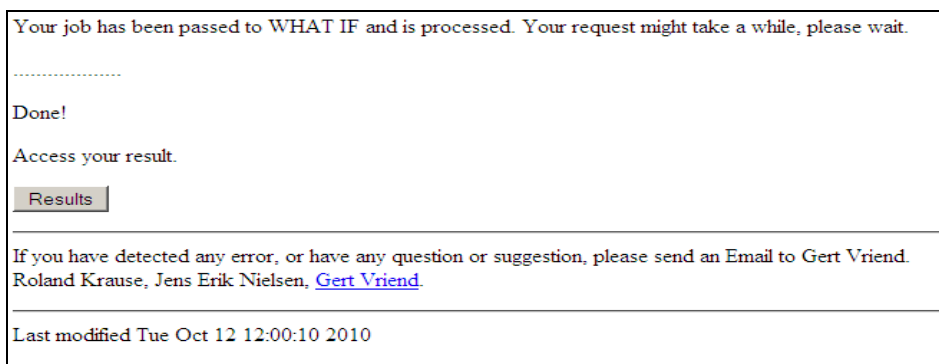


Figure 3.8 Image of job done result on what if weblink

- 4) Selecting the results page (Figure 3.9) a file **predock.pdb** can either be viewed by clicking on the link or saved by saving the link to a file. Rename the file to something sensible e.g. in the case of 2UYN.pdb the file was renamed as 2uynH.pdb

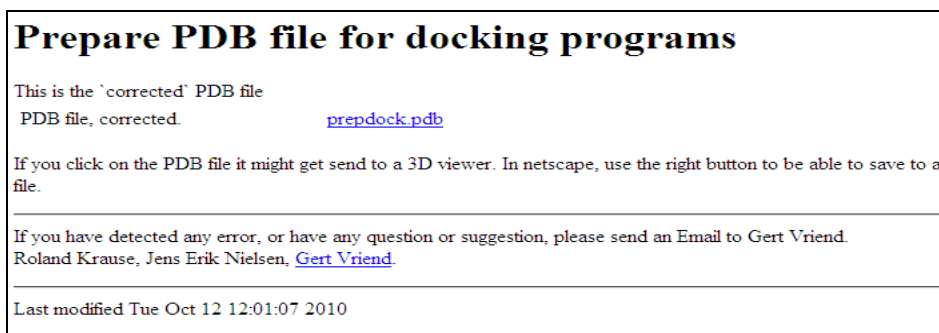


Figure 3.9 what if weblink page image for viewing or saving the .pdb file

### 3.5.2.2 Generating a pharmacophore search model

- 1) Use **Discovery Studio Visualizer (DSV)** by starting the program up on the computer (Figure 3.10).



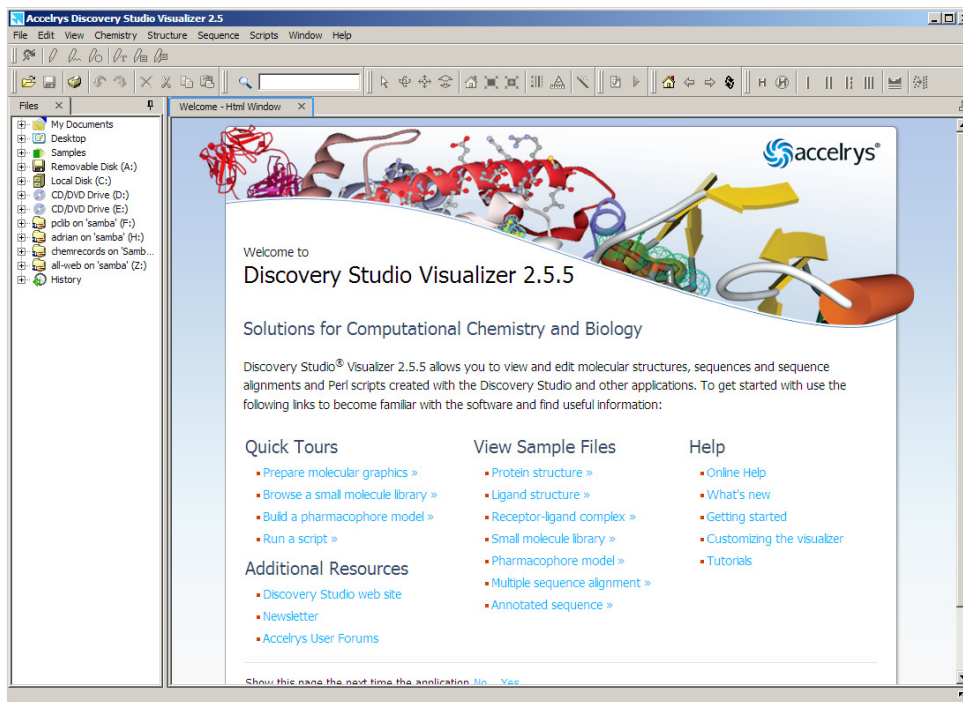


Figure 3.10 Start up image for DSV, used for viewing .pdb file

- 2) Select **File** from the top menu and open the saved PDB file from the respective directory (Figure 3.11).

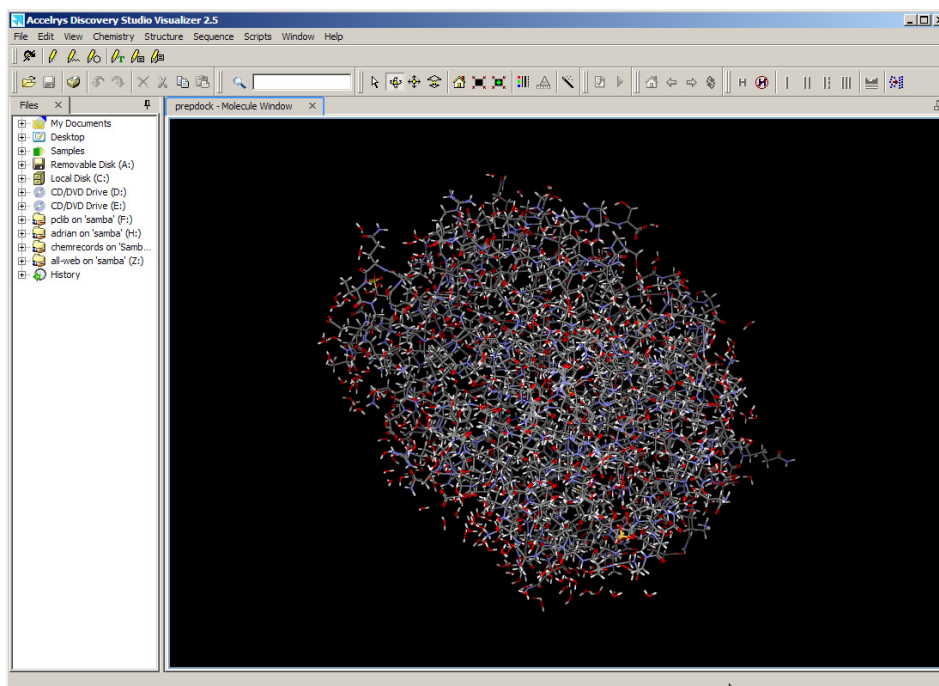

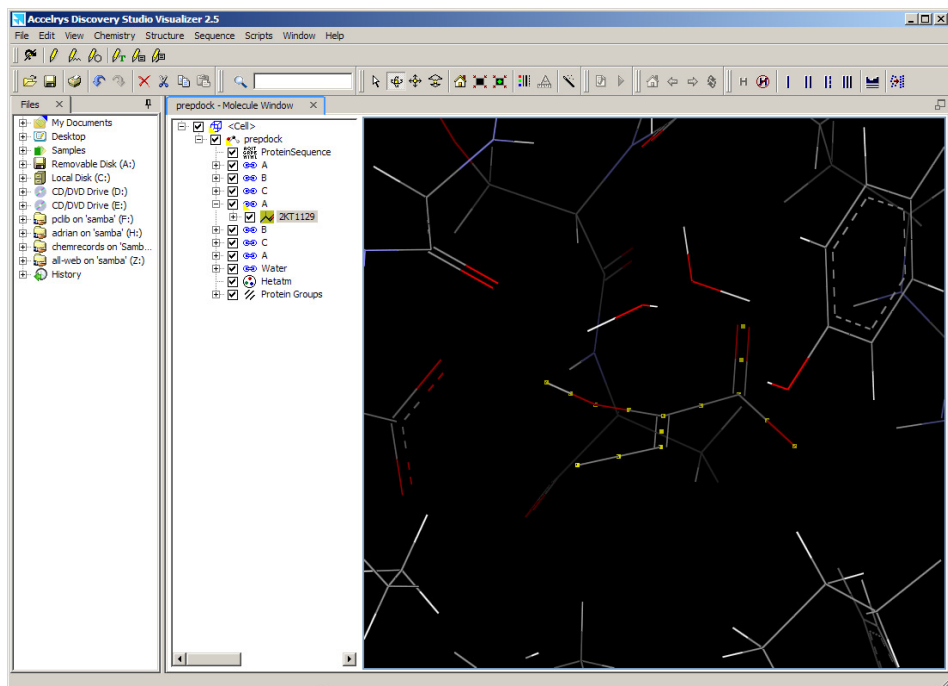



Figure 3.11 3D view of .pdb file in DSV

- 3) A ligand in the active site helps define the pharmacophore. Either way, centre on the active site and select an atom which is roughly at its centre.
  - a. Select **View** from the main menu and then **Hierarchy**, this shows the individual chains and enables to locate the ligand, selecting the ligand in the side menu highlights the ligand structure in the graphics window.
  - b. Click on the  icon in the menu bar to get centered view on the ligand in the active site (Figure 3.12).



**Figure 3.12** Selection of ligand in the binding site of the protein (the selected ligand is highlighted with yellow dots)

- c. If there is no ligand in the binding site, use the graphics window to identify a water atom in the active site cavity and click on the  icon for a centered view.
- 4) To generate the **exclusion spheres** in a radius of about 10Å around the centre of the active site (Figure 3.13).
  - a. Hide the hydrogens at this stage by selecting **Chemistry** then **Hydrogens** then **hide** from the main menu.
  - b. Hide the ligand atoms by selecting from the main menu **Scripts** then **Visualizations** then **Show/Hide Ligands**.
  - c. To select atoms for exclusion spheres select **Edit** from the main menu then **Select**.
  - d. Choose **Radius** and specify 10 for the number of Angstroms and **Apply** and then **Close**.

- e. Check by rotating the graphics window that all amino acids around the active site cavity are selected. The value of 10Å could be made larger or smaller to suit the protein in question.

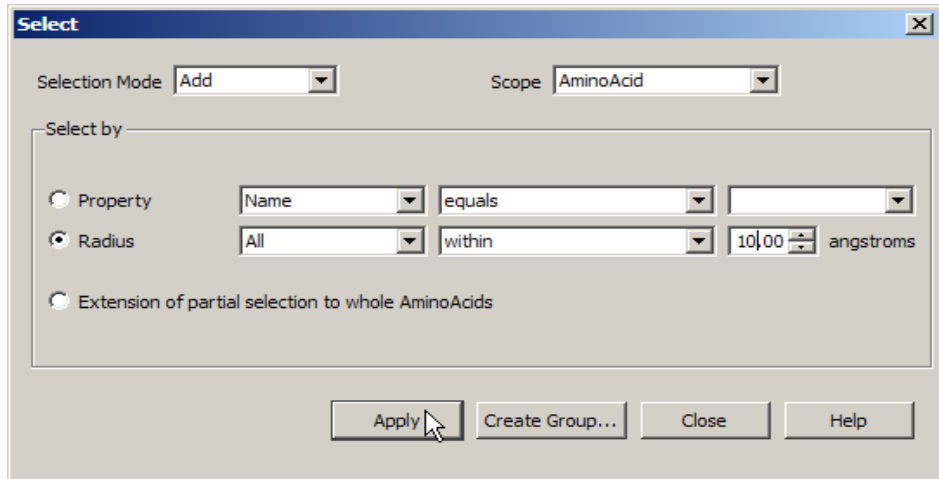


Figure 3.13 Image of window for selection of radius around the binding site in DSV

- f. In case of a mistake use the Ctrl+Z command (as with Microsoft word) to reverse the last action.
- g. Unselect the ligand otherwise exclusion spheres will generate for the ligand too! Do this by holding down the *Ctrl* key and clicking with the left mouse button on the ligand highlighted in the **Hierarchy** window.
- h. With protein atoms still selected (highlighted with yellow dots) select **Structure** from the main menu, then **Query Features**, change the feature from **centroid** to **ExclusionSpheres** (Figure 3.14) in the menu and click ok

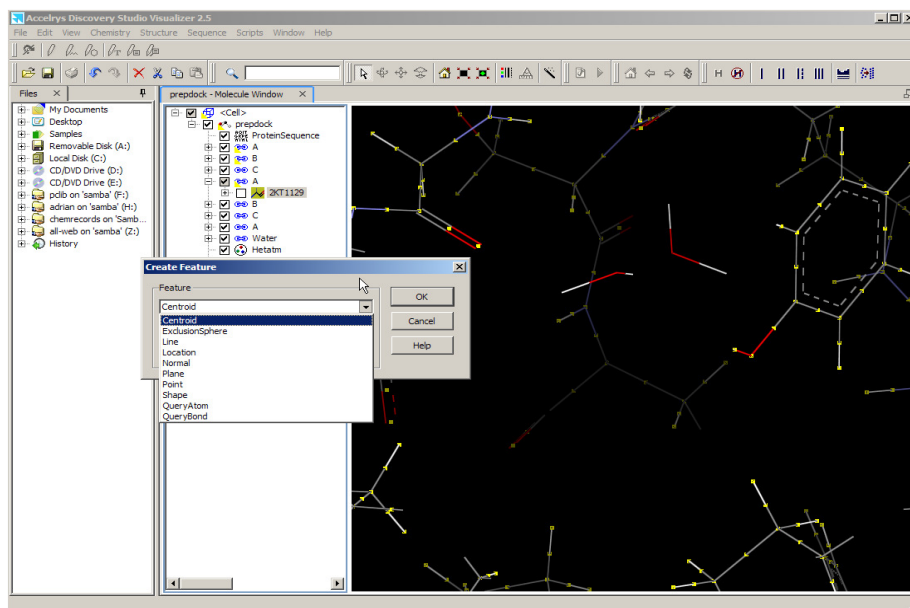


Figure 3.14 Image of DSV with the query feature window

- i. This generates a large number of football-like spheres for the pharmacophore (Figure 3.15). The exclusion spheres are not helpful for visualization so hide them by selecting **View** from the main menu then **Visibility** in the sub menu and then selecting **Hide**.

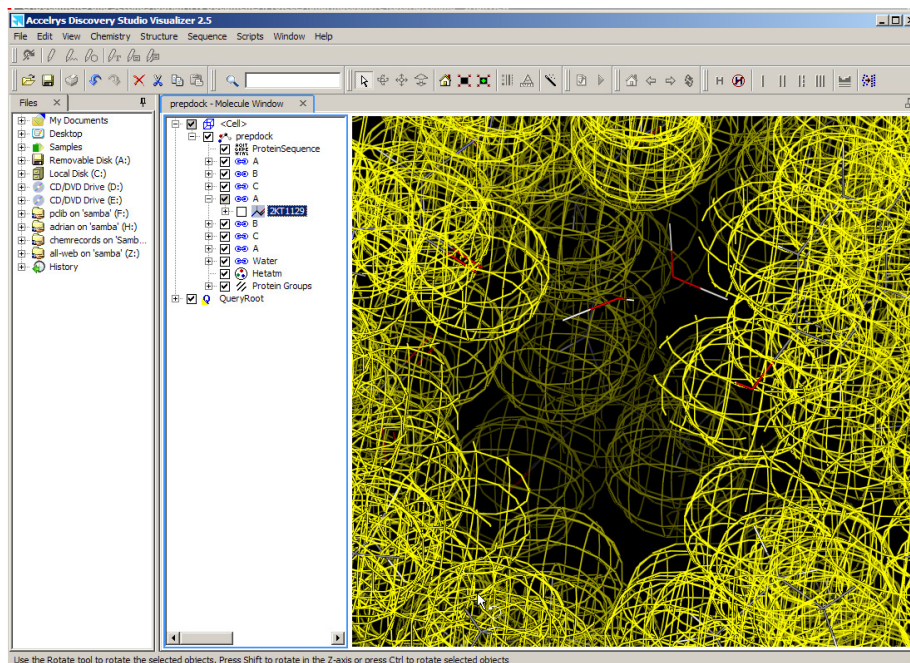


Figure 3.15 Exclusion spheres (highlighted in yellow) around the active site of the protein

### 3.5.2.3 Generation of hydrogen bond acceptor query feature

- 1) To generate H-bond acceptor feature unhide the ligand (either using the **Hierarchy** window or the **Scripts** option on the main menu). In case of no ligand water molecules can be used to generate H-bond acceptor features.
- 2) Select an atom to be a hydrogen bond acceptor on the ligand, typically an oxygen atom (Figure 3.16) and then select **Structure** from the main menu, and then **Query Features** and then **Acceptor**.

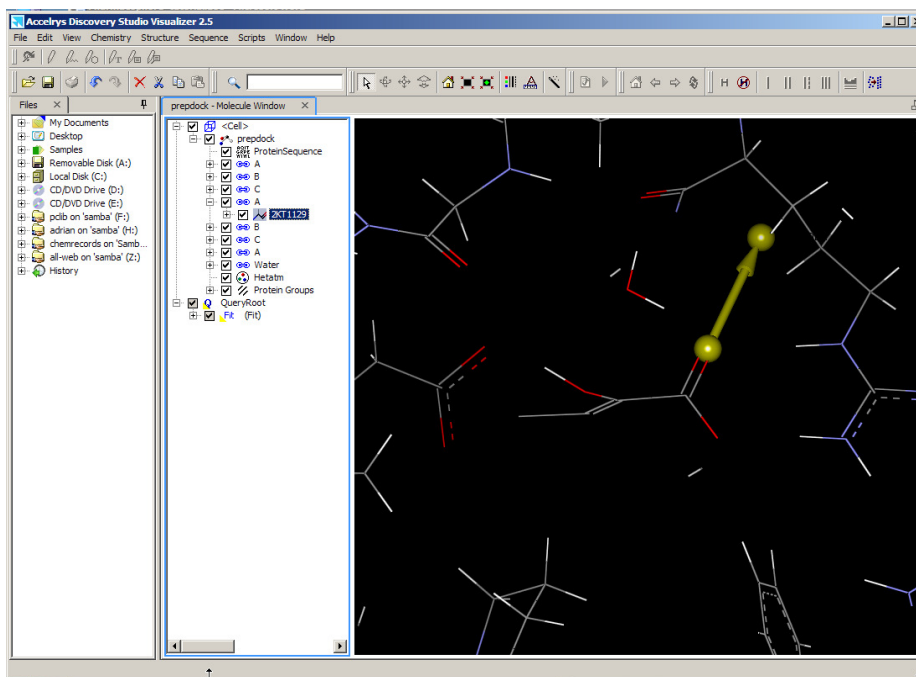


Figure 3.16 Introduction of H-bond acceptor vector from the ligand atom towards amino acid residue of the protein

- 3) The acceptor vector highlighted in yellow is pointing into space and not towards the sensible H-bond donor atom. To change this click the right mouse button on the H-bond acceptor vector and select **Attributes of Acceptor** at the bottom of the menu (Figure 3.17).

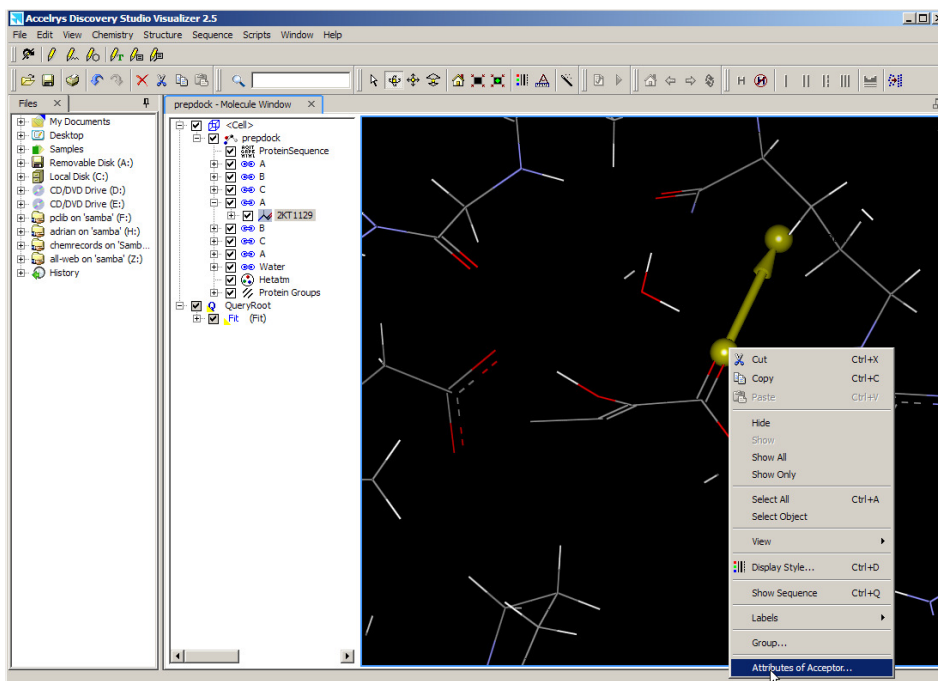


Figure 3.17 Selection window for changing the properties of H-bond acceptor vector in DSV.

- 4) The first thing to change in the **Acceptor feature Attributes** window is the **Orientation** of the vector. Click on the value in the window and it becomes a selectable menu for applying the changes (Figure 3.18). The default is always **Projection**.

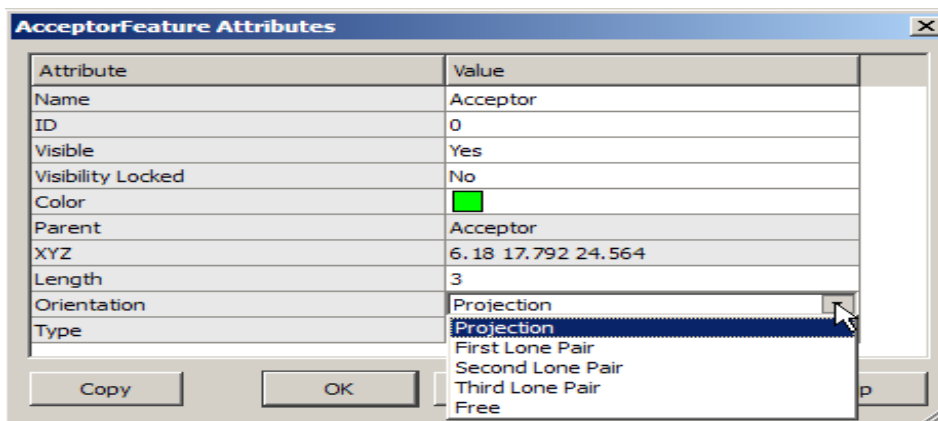
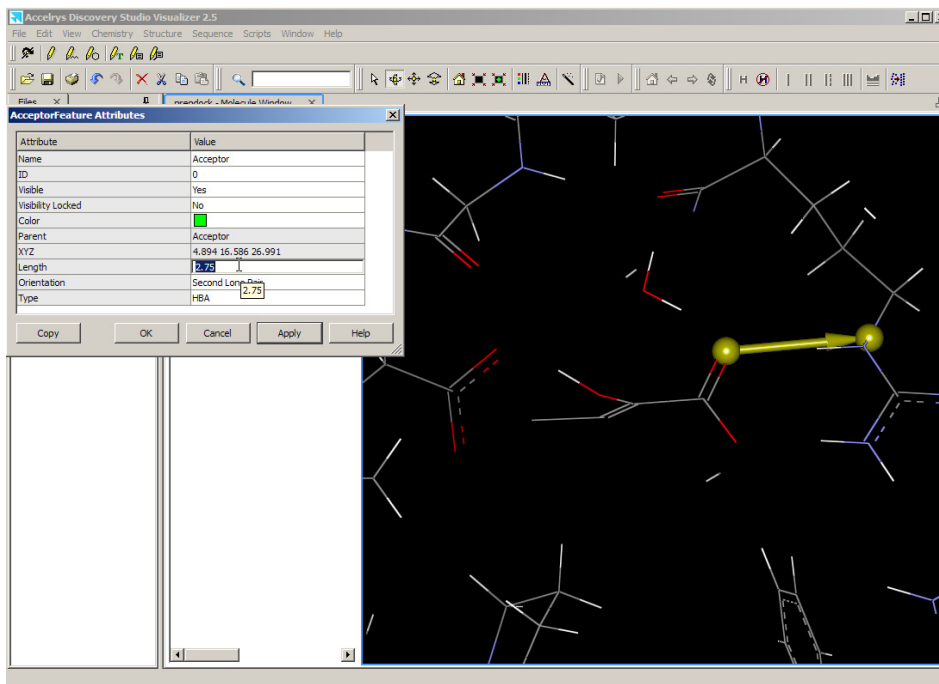


Figure 3.18 Multi options for changing the properties of acceptor feature

- 5) In this case the **Second Lone Pair** gives a sensible vector to the arginine in the diagram. Change the **Length** from 3Å to more sensible value of 2.75Å (Figure 3.19).





**Figure 3.19** Changing the length of the H-bond acceptor vector fits it properly towards the arginine of the protein

- 6) As long as the vector is within a small fraction of an angstrom from the correct position it should be fine. The XYZ values can be changed but this refers to the **Acceptor atom** position which in this case is the ligand therefore no need to change it.
- 7) Same procedure can be used for another acceptor atom by repeating steps 1-6. 2-3 vectors are reasonable for the first pharmacophore.
- 8) The addition of a **Location restraint** is carried out by selecting the *Acceptor Head* points in the graphics window or in the **Hierarchy** window then selecting **Structure** on the main menu then **Query Features** and then from that menu select **Location** (Figure 3.20).

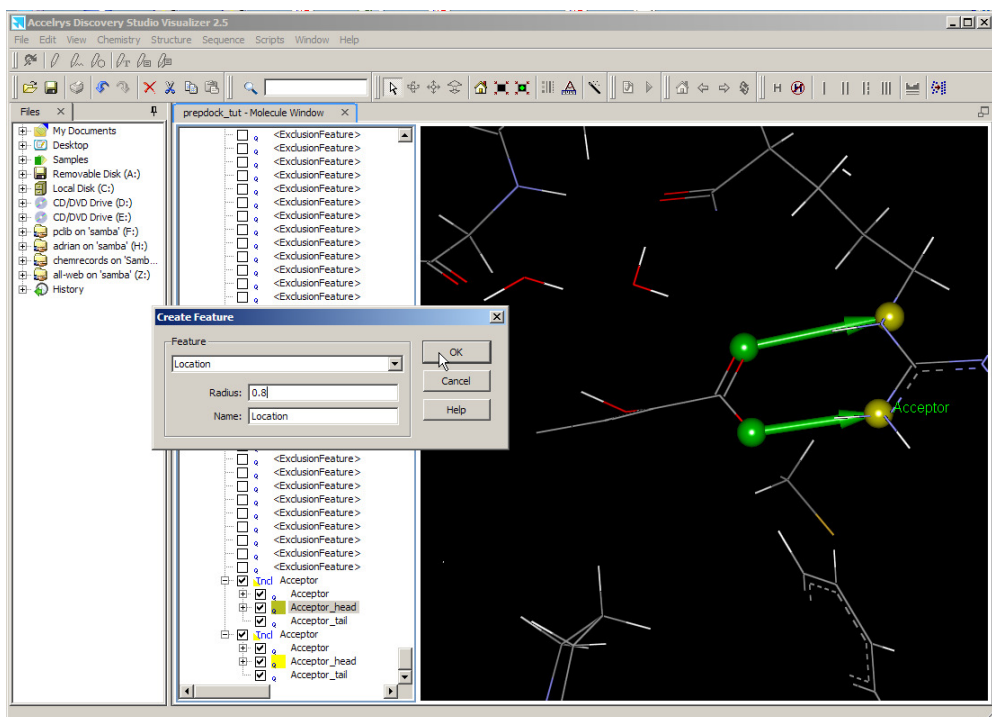


Figure 3.20 Selection of arrow head of H-bond acceptor vector for generating the location restraint

- 9) Due to a bug in the program the radius value for location sphere defaults to 1.5Å. To change this select the *Location spheres* individually and change the **Radius** value manually by hand to 0.8Å for the **Acceptor head Location** and 0.35Å for the **Acceptor tail Location** (Figure 3.21).



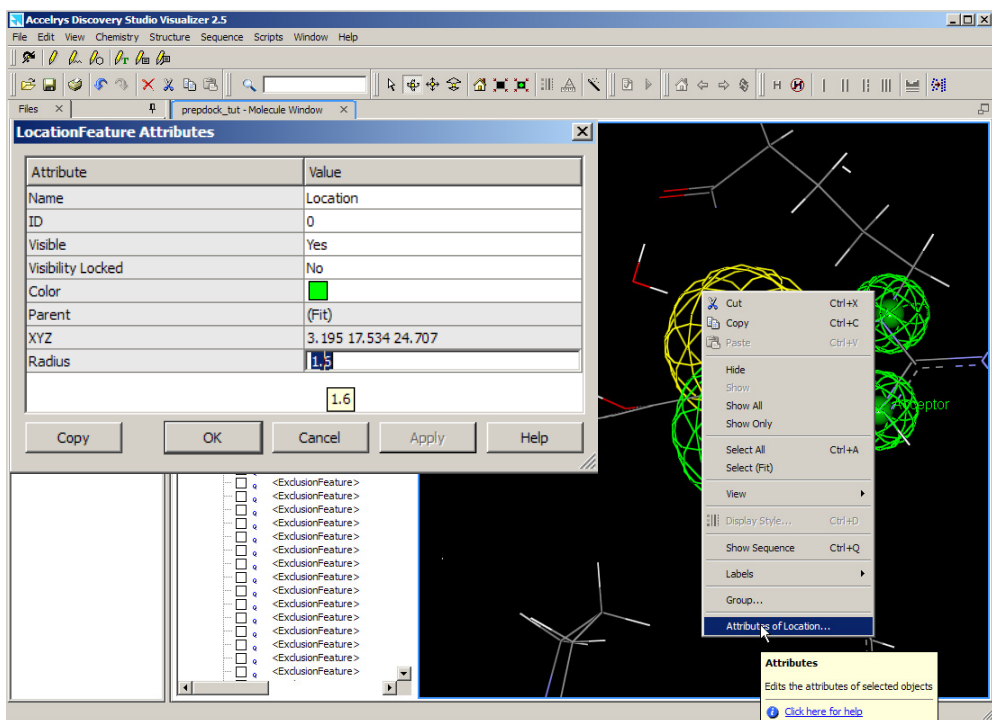
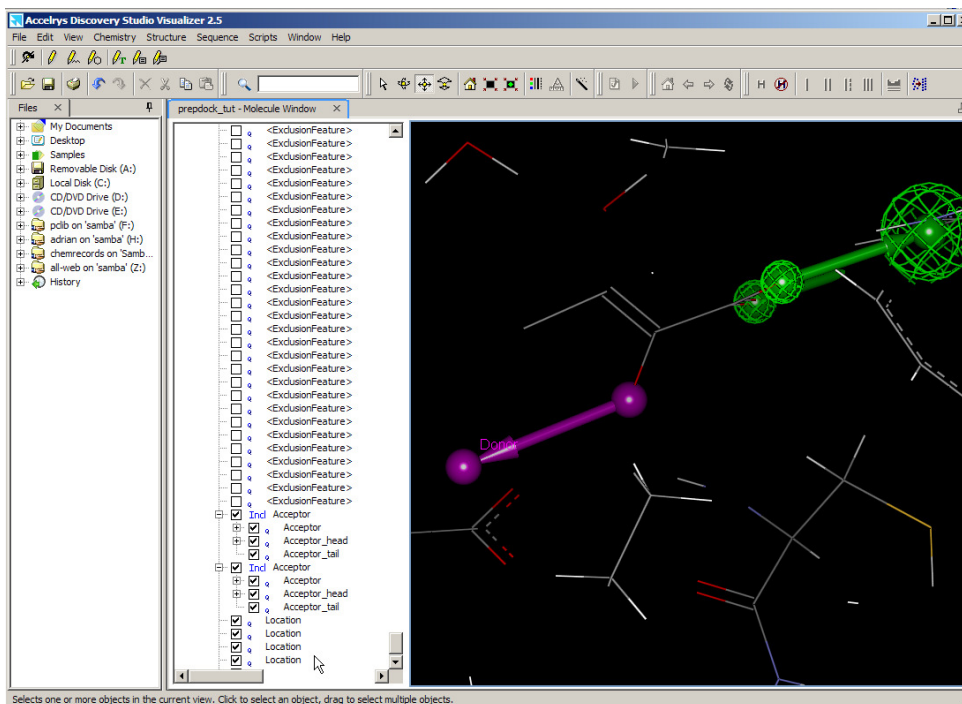



Figure 3.21 Changing the radius of location sphere around acceptor head and tail

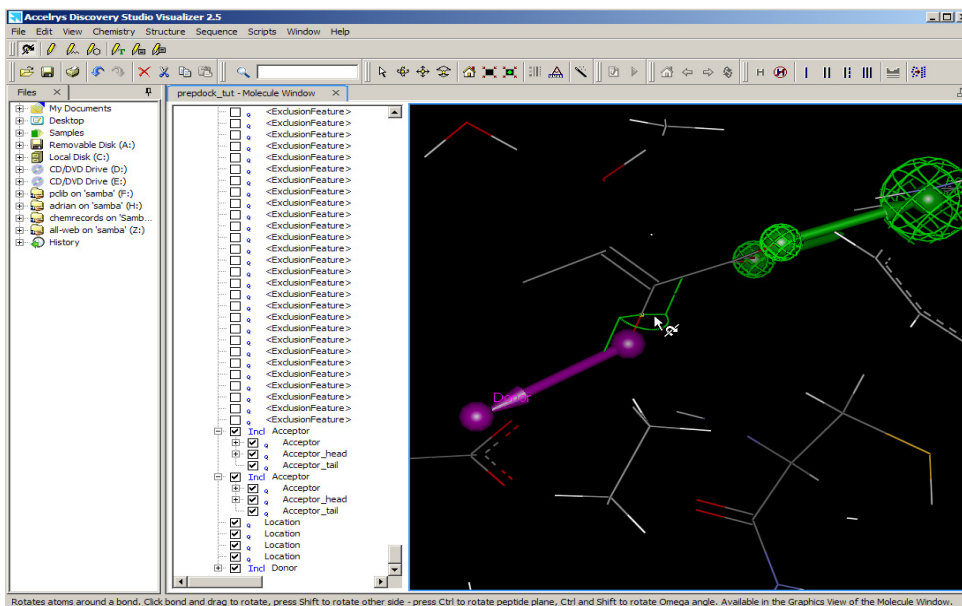
#### 3.5.2.4 Generation of hydrogen bond donor query features

- 1) The H-bond donor feature is generated in the same manner as for H-bond acceptor (section 3.5.2.3) except instead of selecting a hydrogen bond acceptor atom on the ligand, a donor atom such as hydroxyl oxygen or amide nitrogen is chosen.
- 2) The only problem is that the donor hydrogen atom may not point in the correct orientation; instead the **Donor Head Point** could be pointing into space (Figure 3.22). This is a similar problem as for hydroxyl groups as an acceptor point.



**Figure 3.22 H-bond donor from a ligand atom (the donor head is pointing in to the space)**

- 3) It is possible to rotate the donor vector around and hence position it close to a suitable hydrogen bond donor atom. Select the middle of the bond between the carbon and oxygen from which the HB Donor vector originates. Use the torsion option  from the main menu to rotate around the bond and move the donor tail atom to a reasonable position (Figure 3.23).

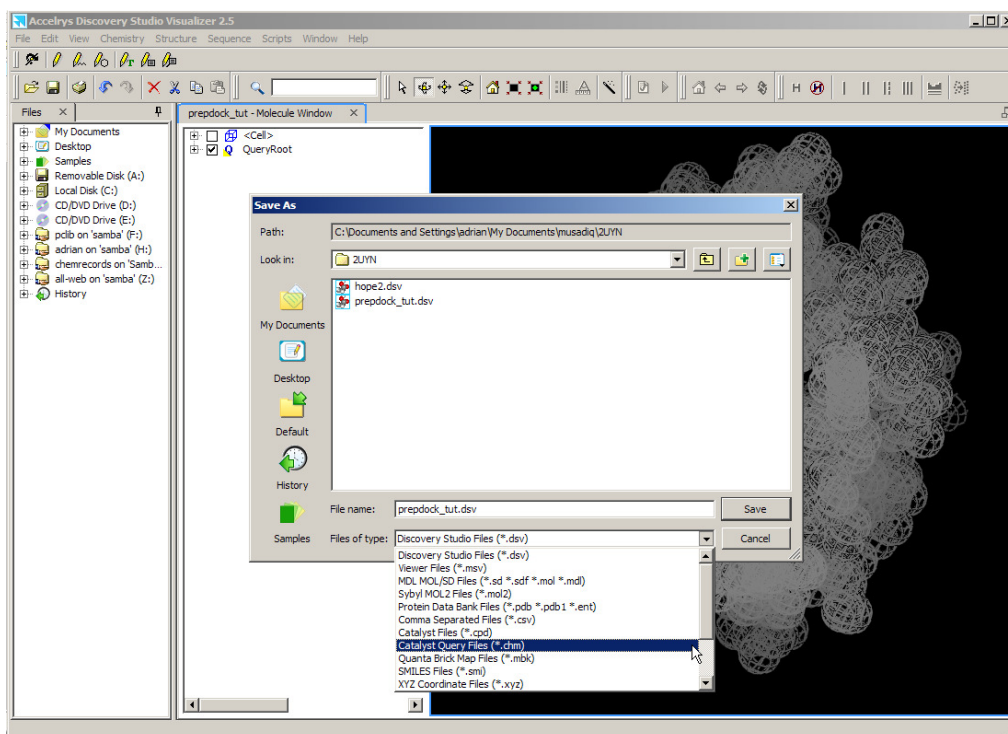


**Figure 3.23 Optimization of the position of the head of H-bond donor vector by using the torsion icon in DSV**

- 4) After optimizing the position of H-bond donor vector add **Location restraints** as described in section 3.5.2.3 (8,9)

### 3.5.2.5 Saving the pharmacophore in .chm file format

- 1) By making all the H-bond vector features visible along with the exclusion spheres, the pharmacophore file can be saved by selecting **File** from the main menu and then **save as** option (Figure 3.24).



**Figure 3.24 Saving the pharmacophore file in .chm file format by using the file, save as option of the main menu of DSV**

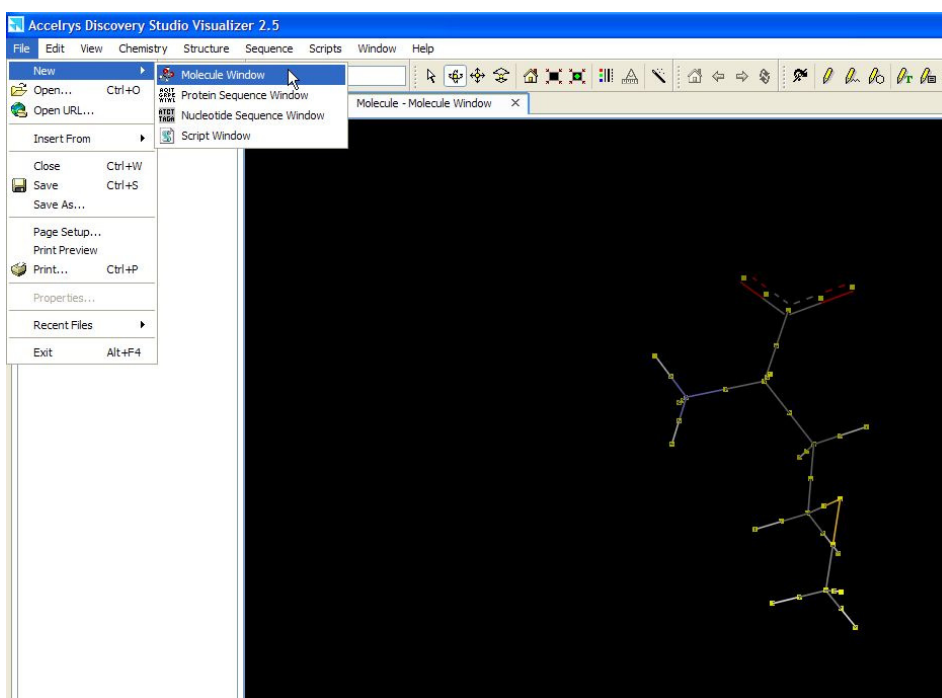
- 2) To search the generated pharmacophore file against the catalyst, the query file needs to be in .chm format. Select a suitable name and use the extension .chm from the **Files of type** option.

### 3.5.3 DSV, Query atom method

Query atom method can be used for the proteins which have the X-ray liganded structure determined, or in which a preliminary vector based approach has identified a fragment/molecule which has chemical features that the user may wish to further explore. In the generation of pharmacophore by using query

atom method, the initial steps involving addition of hydrogens and exclusion spheres are the same as described in vector method section: 3.5.2.1 & 3.5.2.2. After adding Hydrogens to the protein the linked query atoms are introduced with in the binding site of the target protein. Here for pharmacophore generation the protein with PDB code: 1p1m is used. The details of further steps involved are given below:

- 1) First select the ligand as described in vector method 3.5.3.2 (3)
- 2) By simple copying command(ctrl+c), copy the selected ligand and then click on **File** in the menu top and then **New** and then **Molecule Window**, now paste the copied ligand simply by (ctrl+v) command (Figure 3.25).

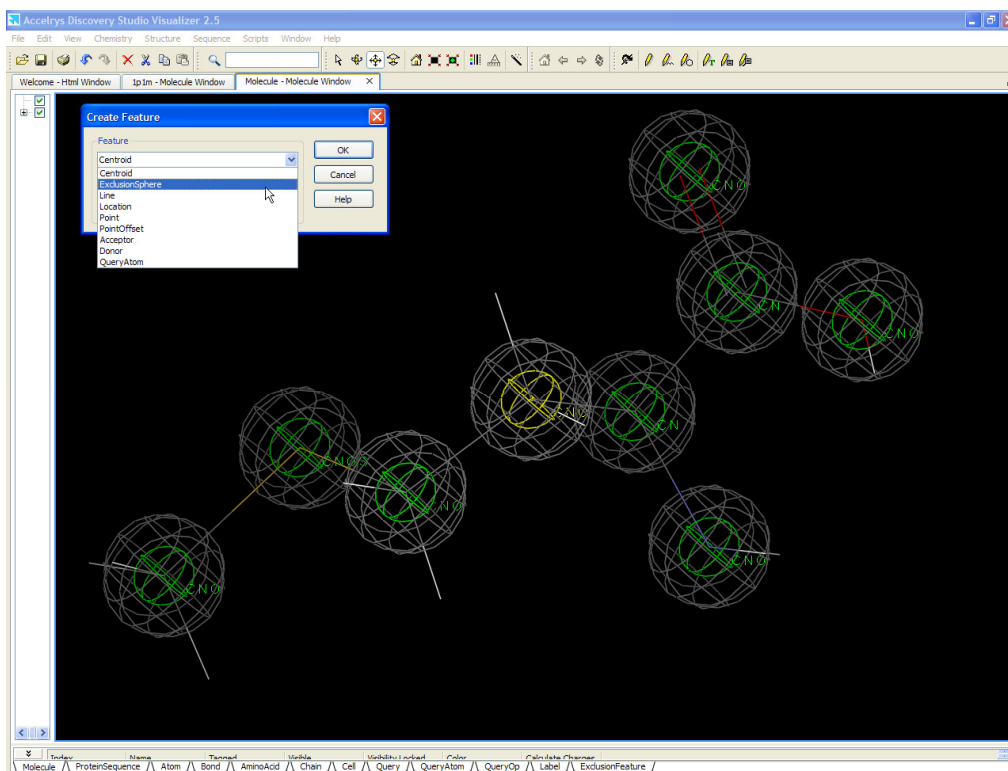


**Figure 3.25 Copying of ligand structure from the pdb file and pasting into new molecule window**

- 3) Select one of the atoms in the ligand and then click on **structure** in the main menu then **query feature** and with in query feature window select **create feature** and then **query atom** option and then in the **List** option to assign the atom type. If the atom type is not known explicitly then a series of potential atom types can be assigned (e.g; O, C, N) and then click ok (Figure 3.26).



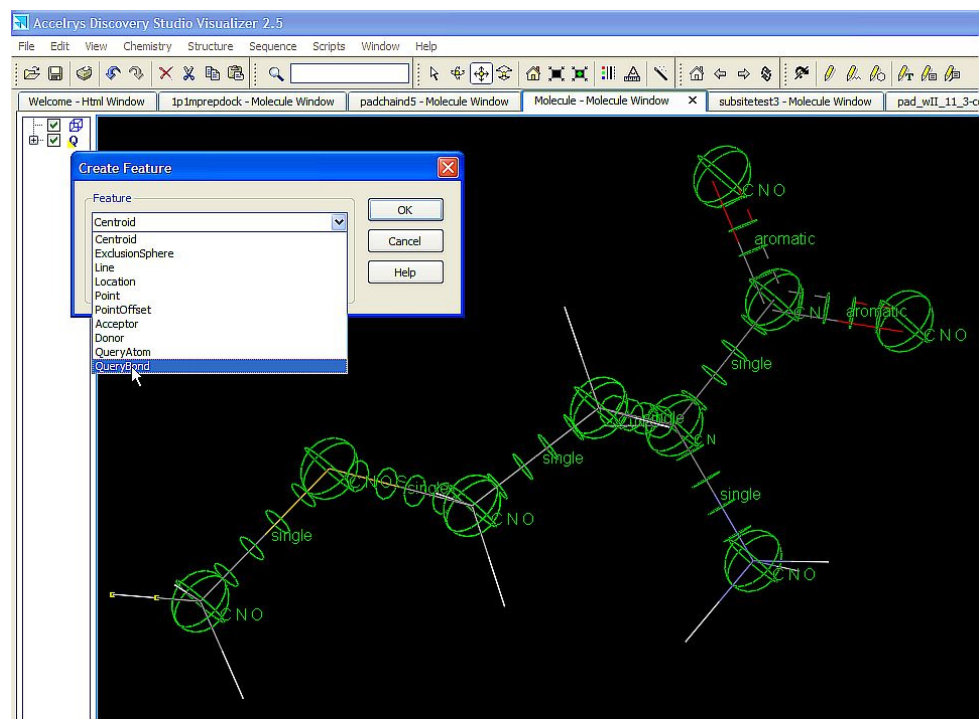
(Figure 3.28). This permits the inclusion of atom linkages within the search with stipulating where the atom must be positioned. By highlighting the location sphere and then right clicking to get to the **Exclusion feature Attributes**, the radius of the location sphere can be increased or decreased accordingly.



**Figure 3.28 Introduction of location sphere (grey) around the individual query atoms by using the create feature of DSV (location sphere have been hidden in the following images for clarity viewing)**

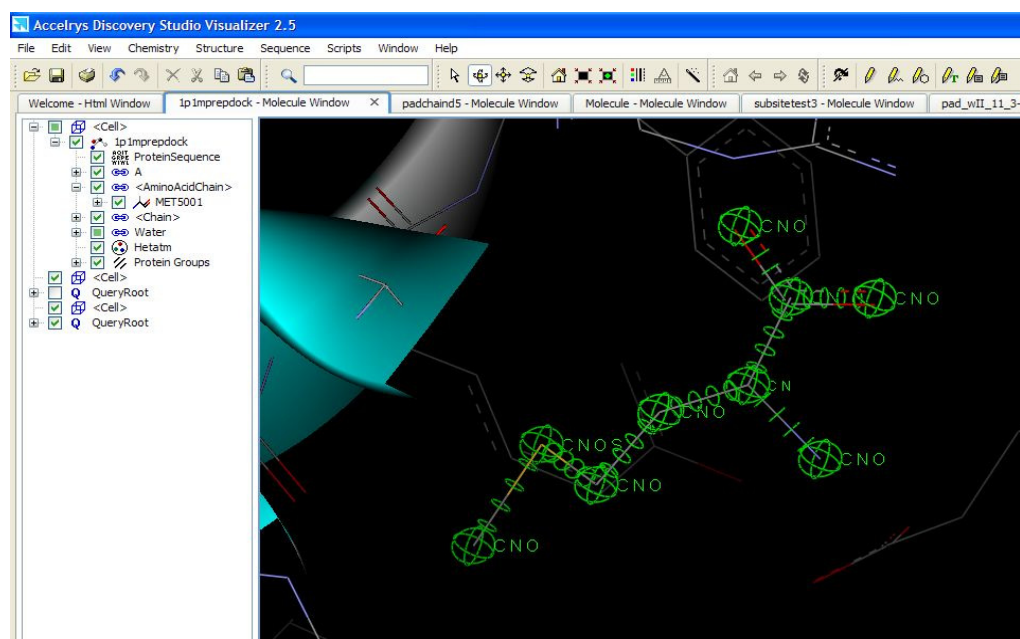
- 6) To introduce bonding between the query atoms, select all the atoms by holding the shift key and click on all the atoms, now click on **structure** in the main menu and select **query feature**, within the query feature select the **query bond** option and click ok. Bonding pattern will appear around the query atoms (Figure 3.29). It is possible to edit the allowed type of bond for any given bond in the same way as carried out for multiple atom types in step 3.





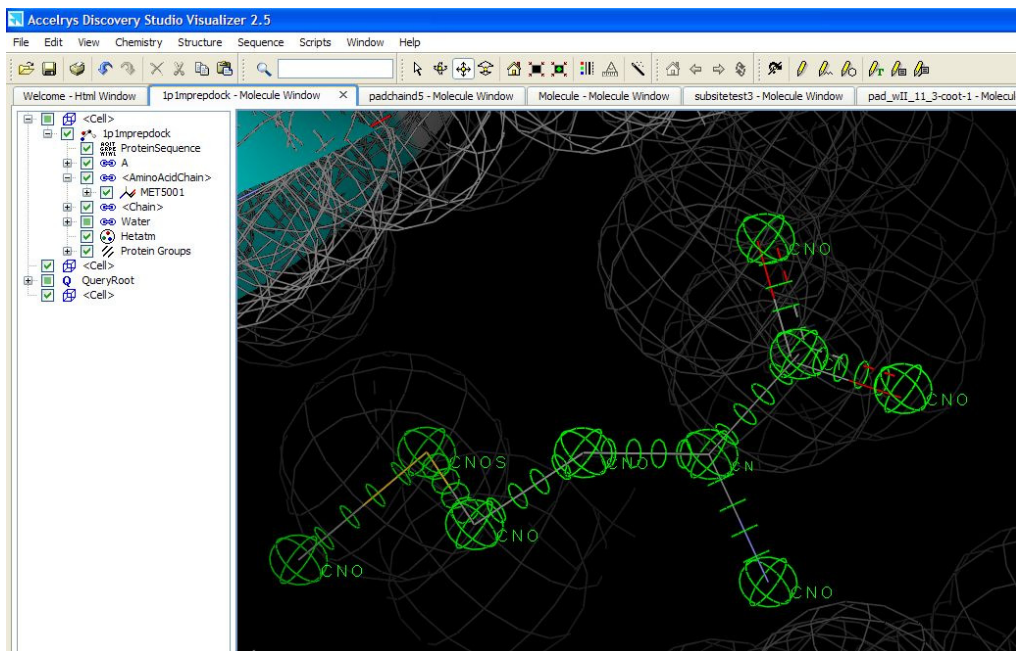
**Figure 3.29** Introduction of bond feature among the query atoms by using the create feature of DSV.

- 7) Now by selecting the whole query copy the whole query ligand and paste it in the protein graphics window of DSV® by using ctrl+a, ctrl+c, ctrl+v commands respectively (Figure 3.30)



**Figure 3.30** Introduction of query ligand in the protein graphics window.

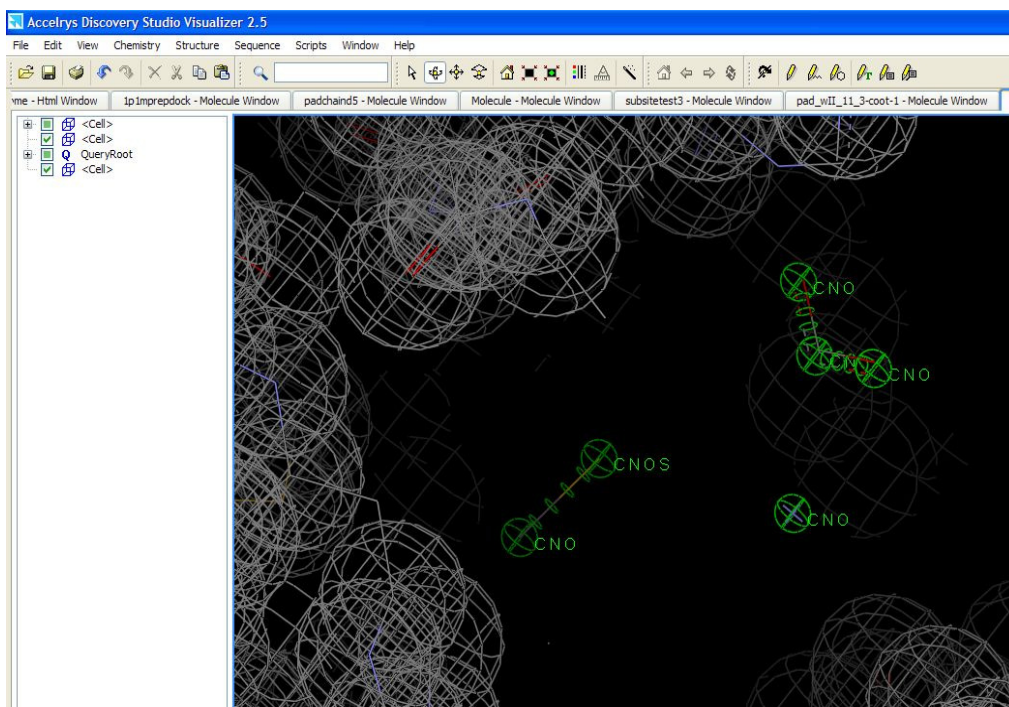
- 8) Introduce the exclusion spheres around the active site of the protein (Figure 3.31) as described in vector method section: 3.5.2.2



**Figure 3.31** view of query ligand along with exclusion spheres in the binding site of the protein.

- 9) Save the file with .chm extension as described in section 3.5.2.5 of vector method, the pharmacophore is now ready to be subjected to database search by using catalyst.
- 10) Interestingly a sub-structure/fragment search can also be carried out by first selecting certain fragments of the query ligand and then deleting the selected fragments by using the delete key of the key board and then saving the file with a different suitable name in .chm file format. Sub structure/fragment based search deletes certain unimportant atoms from the ligand and leaving the important atoms of the ligand in the final query (Figure 3.32). This is an added advantage of the method and has been successfully used in cases of HutD and PARI proteins.





**Figure 3.32** View of a sub-structure/fragment based query ligand in the active site of the protein along with the exclusion spheres.

### 3.6 Advantages of query atom method

The query atom method for pharmacophore generation is comparatively easy and friendly to use in comparison to the other two methods. It allows the user to understand the delicacies of the binding site, to add in sufficient information in the form of geometric constraints and to increase or decrease tolerance values on these geometric constraints. It facilitates the efficient virtual screening and provides ways to understand the chemical interaction between the ligand and the binding site of the target protein. Interestingly it has been used for the first time for this type of study and the results obtained in the form of hits confirm the validity of this type of pharmacophore searching and persuades the user to use this piece of sophisticated software for the following purposes i.e.;

- 1) Using ligand information from a homologous active site architecture to search a new protein structure.

- 2) To specify the important interactions between the potential ligands and the active site of the protein which are almost certain to be retained due to the chemistry carried out by the protein.
- 3) To combine information from multiple ligands and water molecules to create sophisticated searches for potential ligands. This includes the addition of query atoms and vectors if necessary.
- 4) The ability to change the atom position uncertainty spheres, which defines their own uncertainty for certain interactions. This also permits the inclusion of connectivity of atoms without specifying an uncertainty sphere which means the atom type needs to be there but the exact position is unknown.

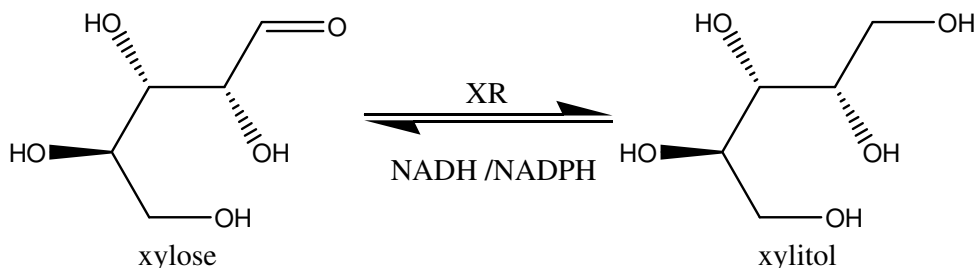
The results obtained from the database search show that some parts of the ligand molecules are essential for interaction and they must assume a particular three dimensional orientation which becomes complementary to the protein active site in order to interact favourably. The methodology provides excellent approach for the optimization of the pharmacophore to attain sensible hits. By following the straight forward procedural steps, we have developed a simple, yet very effective pharmacophore model for the exploration of the binding site of proteins. This not only enables the perception of a pharmacophore within the active site but further through visual pattern recognition also helps the user to research the hits obtained from 3D database searches and to suggest the true substrate/ligand for the protein and thus the probable function of the protein.

This type of pharmacophore modeling search enables the determination of key interactions sites within the three dimensional space of active site, which may have a crucial role in ligand binding. Based on the chemical features of the ligand a tighter or weaker receptor-ligand complex may form. The procedure is automatic and seems quite practical when dealing with a set of different compounds in a large database. The hits obtained help in identifying the key features and constraints which are consistent throughout the optimization of the pharmacophore and vital for fitting the ligand. In short we present a new method for the development of pharmacophore models that accounts both for

the inherent flexibility of the target active site as well as the essential chemical features of the ligand.

## 4. Xylose Reductase

Xylose reductase (XR) is a NADPH or NADH dependent homodimeric oxidoreductase and belongs to the large monomeric aldo-keto reductase (AKR) super family of proteins. It is involved in the assimilation of xylose, where it catalyzes the first step and reduces the open chain xylose in to xylitol (Figure 4.1). The catalytic residues involved in AKR's activity are known and have been studied in the related enzyme aldose reductase{135}. There is a catalytic tetrad comprising of His114, Tyr-52, Asp51 and Lys81 and together they form an oxyanion hole in which the aldehyde or keto group of the substrate binds. The tyrosine acts as a general acid to protonate the oxygen of the corresponding aldehyde or ketone group. In comparison to other AKR's the binding pocket of the enzyme is more polar and is therefore suited for a defined substrate. It has been observed in the crystal structure that the loop which folds over the NADPH co-substrate is slightly disordered in the apo form of the enzyme and becomes more ordered upon binding with the co-substrate in the holo form of the enzyme {136}.

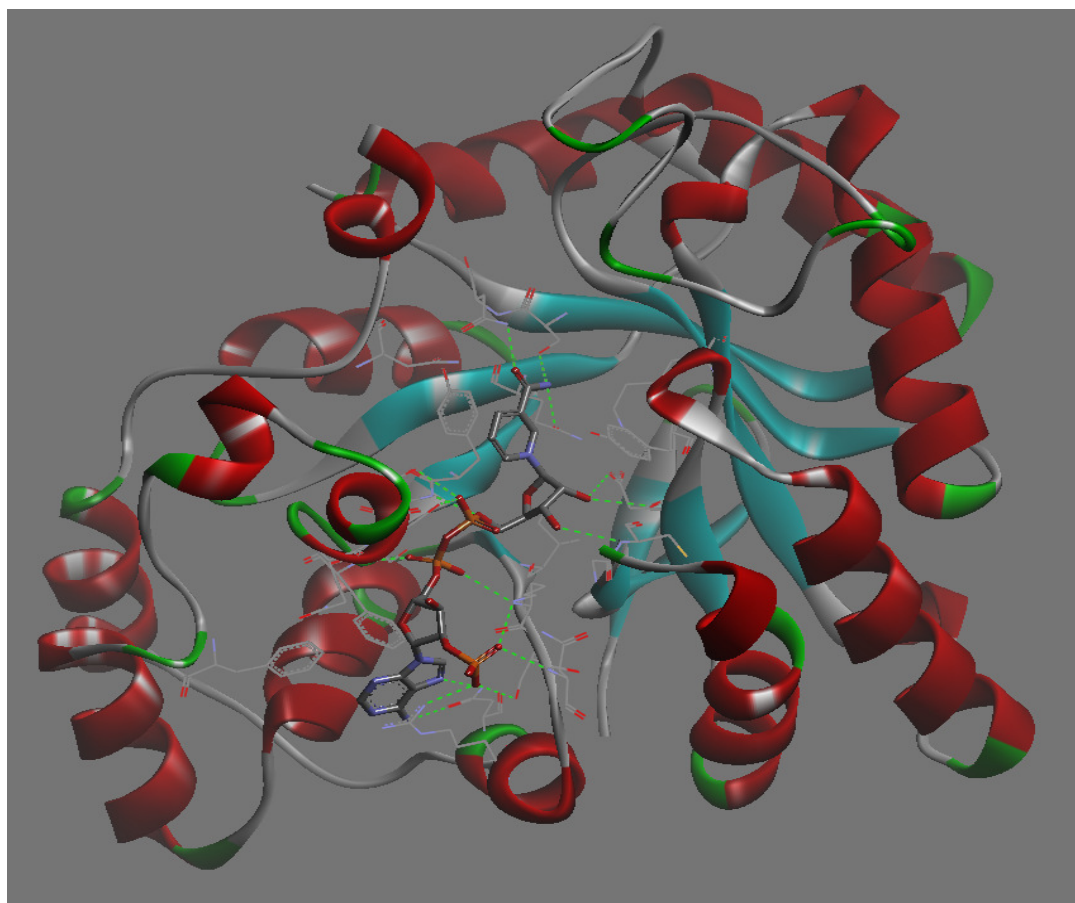


**Figure 4.1 Conversion of xylose to xylitol by XR in the presence of NADH or NADPH**

Clinical research in the AKR's family is principally carried out due to the ability of human aldose reductase to convert the open chain form of glucose to sorbitol. The reaction occurs when the physiological concentration of blood glucose increases in diabetic patients {137}. In order to understand the mechanism and inhibition of aldose reductase by means of different structural studies AKR's have gained pivotal importance in scientific research for more than two decades. Currently no binding mode of the true substrate to any member of the AKR family has been structurally characterized which is probably due to the low concentration of the active form of the substrate and weak binding. In relation to xylose reductases there is scarcity of information about how these enzymes

bind to the substrate? And which interactions are vital among the hydrophobic pocket and the polar substrate?

In order to address these issues we used xylose reductase as a test case to assess the pharmacophore searching method. The objectives were to help understand and to postulate the possible binding mode of various ligands to the enzyme and to see if we can correctly identify all the potential natural substrates of the enzyme that have been identified experimentally. The crystal structures of the two forms of the enzyme (apo and holo) served as good templates for finding the potential ligands/substrates and their probable mode of binding through pharmacophore based searching. Figure 4.2 shows the solid ribbon diagram of Xylose reductase (PDB: 1K8C) Chain : B along with the co-factor NADP forming important H-bonds with the active site residues.



**Figure 4.2 Solid ribbon diagram for Xylose Reductase along with NADP in the active site, forming Hydrogen bonds with the active site residues of the enzyme.**

## 4.1 Pharmacophore searching for xylose reductase

With the availability of co-crystallized structure of xylose reductases {136}, it was possible to locate the potential position of the substrate by using Pharmacophore searching . Additionally there is good experimental data available on a range of different aldehyde and ketone compounds which have been tested and confirmed as potential substrate through enzymatic assays {138}. Although no X-ray structure has been determined of any substrate analogue along with the enzyme, but a potential binding mode has been proposed {136}. There is therefore enough information to check the validity of the method of pharmacophore searching.

The cofactor bound structure of XR (PDB: 1K8C chain:B) was used for the designing various pharmacophores. Factors such as the time of search and the selectivity of a given pharmacophore were noted and the number of features were varied to try and obtain a small set of potential substrates which could be compared to the known substrates. The known substrates were not used to guide the pharmacophore searches as this would defeat the object of this test case.

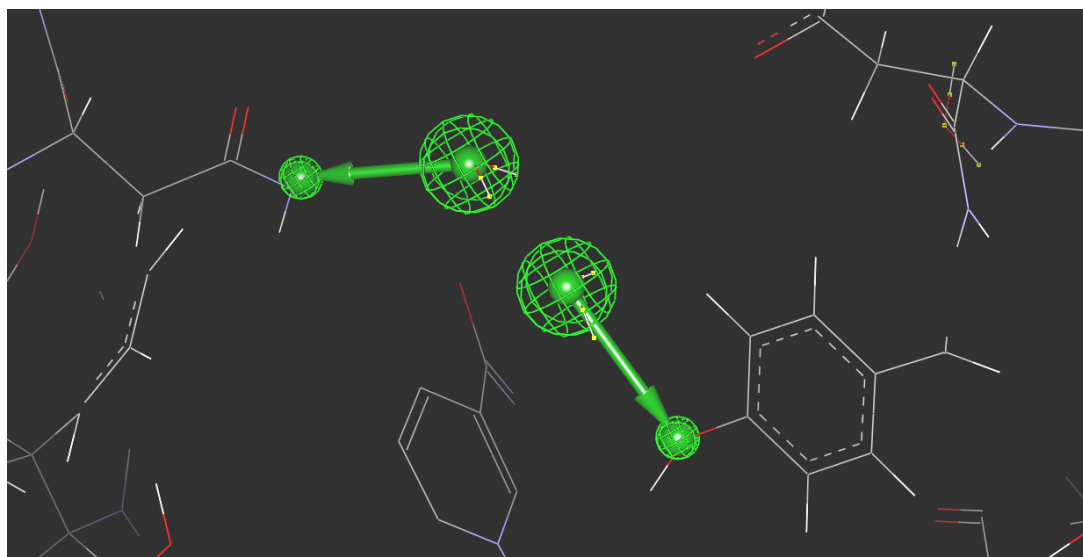
Though the software is not designed for such type of pharmacophore generation. Accelrys® Discovery studio visualizer (DSV), was used for the generation of various pharmacophores. At each stage of designing pharmacophore certain constraints were added for optimization purposes. The most important factors which were introduced from the active site during the course of optimization of the pharmacophore included

- Hydrogen bond donor vector
- Hydrogen bond acceptor vector
- Exclusion Spheres around the active site
- Radius of the area of exclusion spheres around the active site
- Use of certain water molecules in and around the active site
- New database of compounds containing aldehyde/ketone functional group

With the aim to get an optimized pharmacophore, a number of pharmacophore models were generated. The detail account of each pharmacophore model is given below.

### 4.1.1 Pharma xylo 1

In order to generate the pharmacophore the DSV, Vector method was used as described in section 3.5.2. On the basis of information available about the recognition sites and mechanism of aldo/keto reductases. The enzyme has a catalytic tetrad which includes Tyr52, Lys80, Asp51 and His114. Among them Tyr52 and His114 form part of the anion hole in which the aldehyde or ketone group of the substrate binds {139}. A hydrogen bond acceptor vector was added from the hydroxyl of Tyr52 which also corresponds to a water position in the crystal structure (Figure 4.2). Another hydrogen bond acceptor vector was introduced from a water molecule which formed a hydrogen bond with the nitrogen of a conserved Asn310 (Figure 4.2). Further the angle and arrow direction of hydrogen bond acceptor vectors were optimized. The pharmacophore was then search against the small metabolite database (Naturalism.bdb). The hits obtained as a result included compounds like alcohols, aldehydes, ketones, esters, sugars, carboxylic acids, phosphates and aliphatic hydrocarbons.



**Figure 4.3 Pharmacophore with H-bond acceptors pointing from H<sub>2</sub>O molecules towards TYR52 and Asn310 respectively**

**Search time = 15 minutes**

**Maximum limit of Hits = 500**

**Number of Hits obtained = 500**

The type of hits included many compounds other than aldehydes and ketones, which made it clear that a two vector pharmacophore with active site exclusion volumes was insufficiently selective to obtain a suitable subset of hits. In addition the Catalyst software does not allow restriction of a given hydrogen bond acceptor to a subset of chemical groups, namely aldehydes and ketones. As a result a significant number of hits were not aldehydes or ketones. While it is possible to limit the number of hydrogens for a given acceptor site but the ability to restrain the functional groups of the ligands is limited in Catalyst. Given the known enzyme action of this class of enzyme this was not sufficient to restrict the chemical groups in a suitable way. Therefore, a solution was to make a subset of the main library which only contained compounds with aldehyde and ketone functional groups.

#### **4.1.2 *Pharma xylo 2***

##### **Generation of restructured aldehyde/ketone database**

As we know that XR like other aldo/keto reductases works on either aldehydes or ketones it would seem logical to restrict the search to compounds with these functional groups. Therefore a database was created containing compounds with only aldo and keto functional groups as described in section 3.4.1. The numbers of compounds in the new database were reduced to 262 in comparison to the naturalism database having 5,492 compounds. The new database was named as aldoketocomps.bdb. Re-running pharmacophore search gave the following results.

**Search time= 31 minutes**

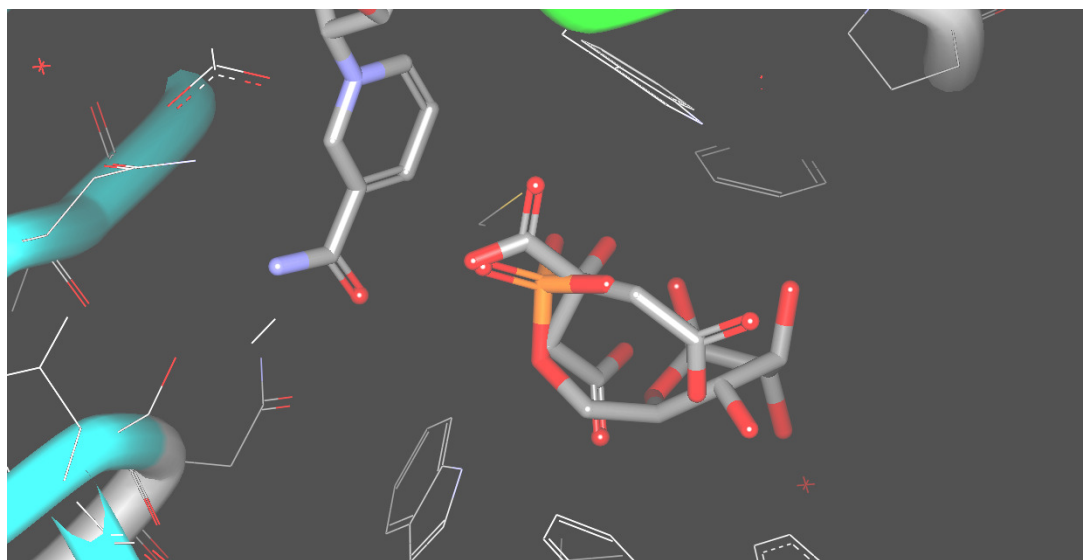
**Maximum limit of Hits = 262**

**Number of Hits obtained = 71**

In the above pharmacophore the radius of the uncertainty spheres around the head and tail of the vectors was 0.35 and 0.8Å respectively, meaning that there was greater uncertainty in the position of the ligand acceptor position than the protein donor atoms. Out of 71 hits 43 compounds were aldehyde and 3 were ketone consistent with their functional group interacting in the anion hole of the enzyme. The remaining hits had carboxylic acid or phosphate groups towards this position. The hits containing carboxylic acids and phosphate groups in the active



site and to their justification that they can be accommodated into this position in aldo/keto reductases is confirmed by the crystal structures of aldose reductase in complex with citrate (PDB: 2ACS) {140} and glucose phosphate (PDB:2ACQ) {140} (Figure 4.3)



**Figure 4.4 Superimposed X-ray structure of Citrate and Glucose-6-Phosphate in the active site of Xylose reductase with carboxyl and phosphate group towards the oxy-anion hole (the nicotinamide ring,citrate and glucose-6-phosphate represented as stick model)**

As we have observed before, tighter restrictions on uncertainty spheres around the pharmacophore tail (ligand acceptor position) can give better discrimination. Therefore the radius of spheres around the head and tail of both the H-bond acceptor vectors was reversed i.e; 0.8Å and 0.35Å respectively. This is the opposite of what might be expected as the protein positions are well defined but the ligand position is more uncertain. However the uncertainty on the protein position is required to accommodate certain angular restrictions. As expected the Pharmacophore search gave smaller number of hits as shown below.

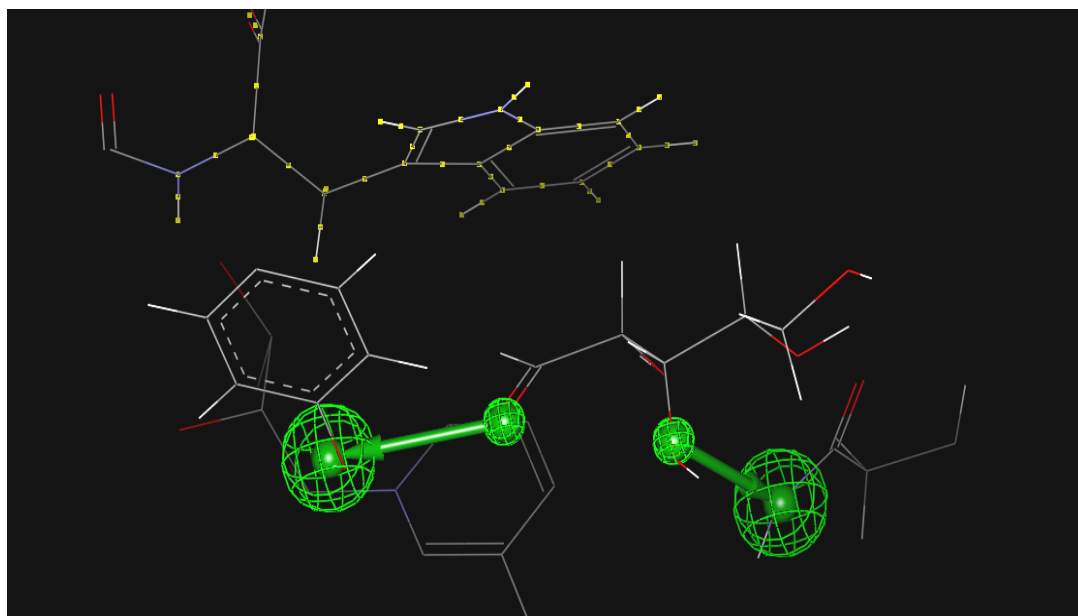
**Search time = 22 minutes**

**Maximum limit of Hits = 262**

**Number of Hits obtained = 54**

Out of the 54 hits, 30 compounds were aldehyde and one was ketone with their functional groups interacting in the anion hole of the enzyme. In both search results there was a clear preference of aldehyde compounds over ketones given the pharmacophore included two hydrogen bond acceptors and exclusion spheres within a radius of 10Å. Analysis of the structure of the active site of XR showed that the side chain of Trp24 restricts the space available around the anion hole

of the enzyme and the presence of an aliphatic group in this area could cause steric hindrance towards Trp24. As a result ketone containing compounds cannot satisfy the pharmacophore and thus probably were not selected among the hits. The proximity of Trp24 side chain towards the aldehyde of D-Xylose (Figure 4.4) shows that presence of ketonic group could cause steric hindrance towards Trp24.

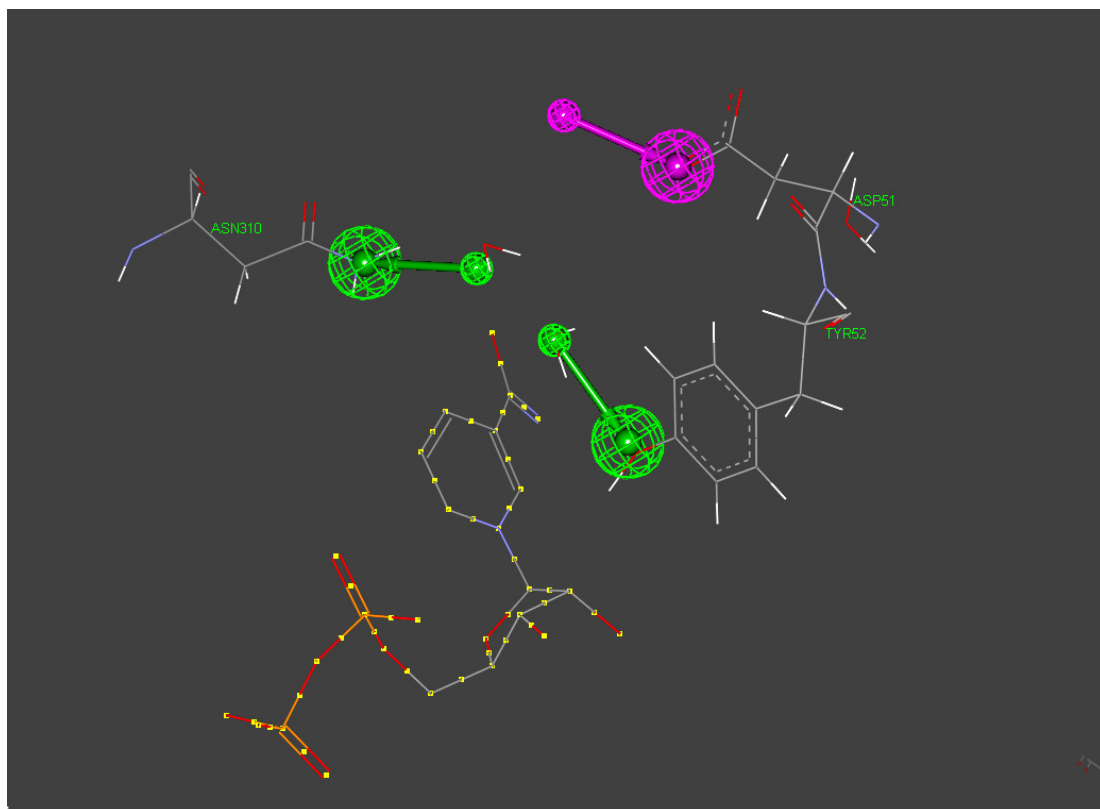


**Figure 4.5** The possible clash of a ketone with the Trp24 (structure highlighted with yellow dots) in the active site(in figure the ligand is D-Xylose)

### 4.1.3 Pharma xylo 3

By observing the active site of XR along with the H-bond interactions made by the hits obtained it appeared that there could be further interactions within the active site other than the specified two H-bond acceptors interactions. A thorough examination of the active site revealed for example the key residues which have been identified for H-bond interactions include Trp24, Asp51, Tyr52 and Asn310. This has also been identified in the X-ray structure of XR {136} that these residues may facilitate the binding of substrate within the active site. In order to further test the capacity of pharmacophore searching in selecting suitable ligands as hits an additional Hydrogen bond donor vector was introduced from the free carboxyl end of Asp51 (Figure 4.5). The arrow head of all the vectors were set on the grounds of previous hits able to form H-bonds. Further

the angle and direction of the vector were optimized. The pharmacophore was then subjected to search against aldoketocomps.bdb database.



**Figure 4.6 Pharmacophore with additional H-bond donor (in magenta) towards the carboxyl oxygen of ASP51, NADP<sup>+</sup> is highlighted with yellow dots and H-bond (exclusion spheres are removed for clarity).**

**Search time = 06 minutes**

**Maximum Number of Hits = 262**

**Number of Hits obtained = 50**

Among the hits 44 compounds were aldehydes 1 was a ketone and the rest were carboxylic acids and esters. As expected the addition of another constraint in the form of a hydrogen bond donor reduced the number of hits although not excessively and showed that it improved the pharmacophore in the sense of getting hits with maximum number of compounds likely to be the substrate/ligand. In addition the orientation of the best hits from the screening matched with D-galactose, where the C2 is oriented towards the cofactor (NADPH) in such a way, which is consistent with hydride transfer from the cofactor NADPH.

#### 4.1.4 Pharma xylo 4

On the basis of previous knowledge available of potential interactions of key amino acids {136} within the active site the final optimized pharmacophore was generated by putting an additional Hydrogen bond donor vector from the ASN 310 (Figure 4.6). Among the hits one of the convincing hits obtained as a result was D-galactose, which was fitting well in the active site by satisfying maximum number of interactions, hydroxyl group of C2 and C3 of D-galactose formed hydrogen bonds with the amino nitrogen and carbonyl oxygen of Asn 310 respectively and its aldehydic functional group was at a reasonable distance from the cofactor for hydride transfer. The final pharmacophore seemed to pose reasonable interaction, as most of the hits obtained as a result of this pharmacophore had their functional group positioned towards the cofactor. This could allows the hydride transfer to occur from the cofactor to substrate, subsequently leading to the proton transfer from the Tyr52 onto the substrate to convert it to its corresponding reduced form as described in detail by {141}. (Figure 4.6)

Searching time = 02 minutes

Maximum Number of Hits = 262

Number of Hits obtained = 51

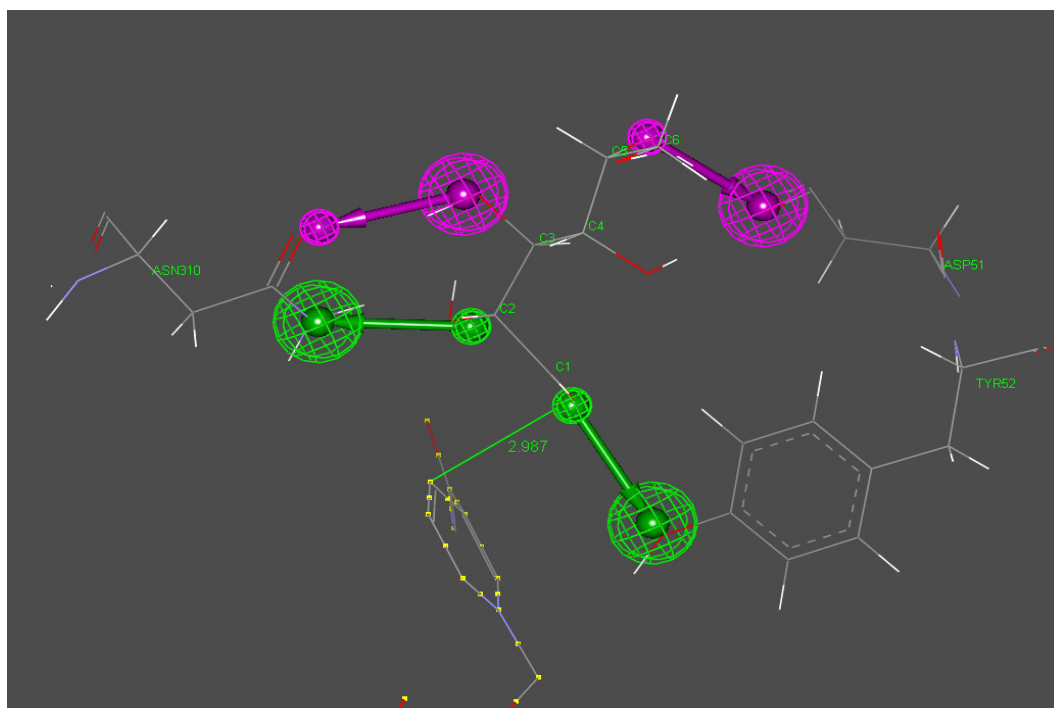


Figure 4.7 Pharmacophore with additional H-bond donor(in magenta) towards the carbonyl oxygen of Asn310,D-galactose forming H-bonds with ASN310 and the position of its aldehydic group at 2.9Å distance from cofactor NADP<sup>+</sup> (highlighted with yellow dots)

#### ***4.1.5 Manual selection of hits***

Among the hits D-galactose seemed to be a convincing hit as the orientation of its aldehydic group towards the cofactor is similar to the suggested mechanism for hydride transfer {141}. By keeping the orientation aldehyde group of D-galactose around the cofactor as standard, the 51 hits from final pharmacophore model and some previous pharmacophore were observed manually. Only 25 compounds were found to have their aldehyde group positioned towards the cofactor for the correct transfer of hydride.

Table 4.1 includes the hits conforming to the D-galactose, among the hits some have already been tested as a substrate on aldose reductase and have demonstrated reasonable activity as a substrate e.g; D-galactose, D-ribose, L-arabinose, D-xylose and D-glucose. Compounds which are among the hits but have not been tested as substrate can be tried invitro to test their catalytic activity. On the basis of results obtained, xylose reductase a type of enzyme which requires a cofactor for it's functioning, appears to be an excellent test case example for the application of pharmacophore searching method and gives the plausible clue of substrate mode of binding and mechanistic details of structure-function relationship on molecular level. This example supports the purpose of using pharmacophore searching, and shows that this method works.

**Table 4.1 Results of hits obtained from final pharmacophore search among the hits are those compounds which align with D-galactose and those which have already been confirmed as substrate for aldose reductases**

S.No	Hits Obtained from final Pharmacophore Search	Hits Aligning with D-galactose	Hits previously confirmed as xylose reductase substrate	S.No	Hits Obtained from final Pharmacophore Search	Hits Aligning with D-galactose	Hits previously confirmed as xylose reductase substrate
1	D-allose	✓		27	D-altrose		
2	D-galactose	✓	✓	28	D-quinovose	✓	
3	D-galacturonate	✓		29	L-lyxose		
4	D-ribose	✓	✓	30	rutinose	✓	
5	D-fucose	✓		31	D-gulose		
6	L-arabinose	✓	✓	32	D-talose		
7	D-aldohexoses			33	D-rhamnose	✓	
8	D-glucuronic acid			34	D-threose		
9	L-galactose			35	abequose	✓	
10	D-apiose			36	D-idose		
11	D-fructuronate			37	isomaltose		
12	(4S,5S)-4,5-dihydroxy-2,6-dioxohexanoic acid			38	L-iduronic acid		
13	D-lyxose			39	L-guluronic acid		
14	D-xylose	✓	✓	40	digitalose		
15	L-xylulose 5-phosphate			41	D-glucuronate	✓	
16	D-arabinose 5-phosphate			42	2-N,6-O-disulfo-D-glucosamine		
17	D-mannose			43	digitoxose		
18	D-allose 6-phosphate			44	ascarylose	✓	
19	D-glucose 1,6-bisphosphate			45	tyvelose		
20	5-dehydro-4-deoxy-D-glucuronic acid			46	D-mannuronate		
21	L-fucose			47	boivinose		
22	D-glucose	✓	✓	48	melibiose		
23	D-galacturonic acid			49	5-dehydro-4-deoxy-D-glucuronate		
24	L-xylose			50	L-iduronate		
25	L-threose	✓		51	L-gulonate		
26	colitose						

## 4.2 Conclusions based on search models and hits

- 1) The enzymatic assays previously carried out by Neuhauser *et al* {138} gives a comparative account of activity of the enzyme against various substrate compounds belonging to aldehyde (Table 4.2). The hits such as D-aldohehexoses are generic and represent a flexible subset of C6-aldehydes and so can be looked at further.

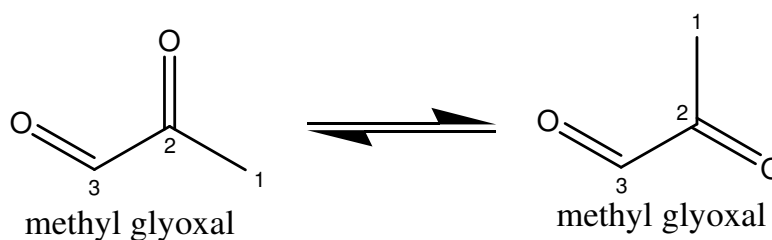
Aldehyde	Relative activity (relative catalytic efficiency)
D,L-Glyceraldehyde (20 mM)	0.9 (37)
D-Erythrose (20 mM)	1.8 (100)
D-Xylose (300 mM)	1 (1)
L-Arabinose (300 mM)	0.7 (2)
D-Ribose (300 mM)	0.3 (0.2)
D-Glucose (300 mM)	0.2 (<0.1)
D-Galactose (300 mM)	0.7 (0.3)
D-Mannoheptose (240 mM)	<0.1 (<0.1)
D-Xylosone (25 mM)	1.1 (20)
D-Glucosone (25 mM)	1.1 (22)
Methylglyoxal (50 mM)	1.4 (20)
Phenylglyoxal (12 mM)	0.6 (17)
Valeraldehyde (50 mM)	0.8 (13)
Pyridine-2-aldehyde (10 mM)	1.0 (7)

**Table 4.2 List of compounds belonging to aldehyde group Identified as substrates with comparison of their relative activity, reactions in aldehyde reduction performed with 17.4 nM ALR at 25 °C in 300 mM sodium phosphate buffer, pH 7.0, with saturating (relative activities) and non-saturating substrate concentrations (1 mM; relative catalytic efficiencies). The constant coenzyme concentration was 250 µM NADH, table reproduced from { 138}**

- 2) Some of the substrates like D-Xylosone and D-Glucosone have not been search out as hits from the database because in the catalyst software the stereochemistry is not defined, resulting in more degrees of freedom which leads to a potential possibility that the conformational sampling of the compounds is compromised, and thus penalizing the outcome of search at the cost of missing potential hits. When the stereochemistry of the D-aldohehexoses was specified by using the Catalyst software and then

subjecting it to search through the naturalism database to search for compounds with the same stereochemistry as D-aldohexoses the search only gave D-galactose as a hit. Thereby further narrowing down the search in terms of specific stereochemistry.

- 3) Compounds like methyl glyoxal, D-xylosone and D-Glucosone which are identified as true substrates {138} and are also present in the database but are not detected among hits. This could be because in the database there are only two conformations possible for the aldo-keto group (Figure 4.7). The bond length between C2-C3 is  $1.54\text{\AA}$ , which is a single bond length which means that no partial double bond lies there. The software logic statement may presume it to be a partial double bond and thus searching for the corresponding conformations or on the other side there could be a bug in the program which rules out the possibility of methyl glyoxal and its derivative aldo-keto compounds to be potential hits.



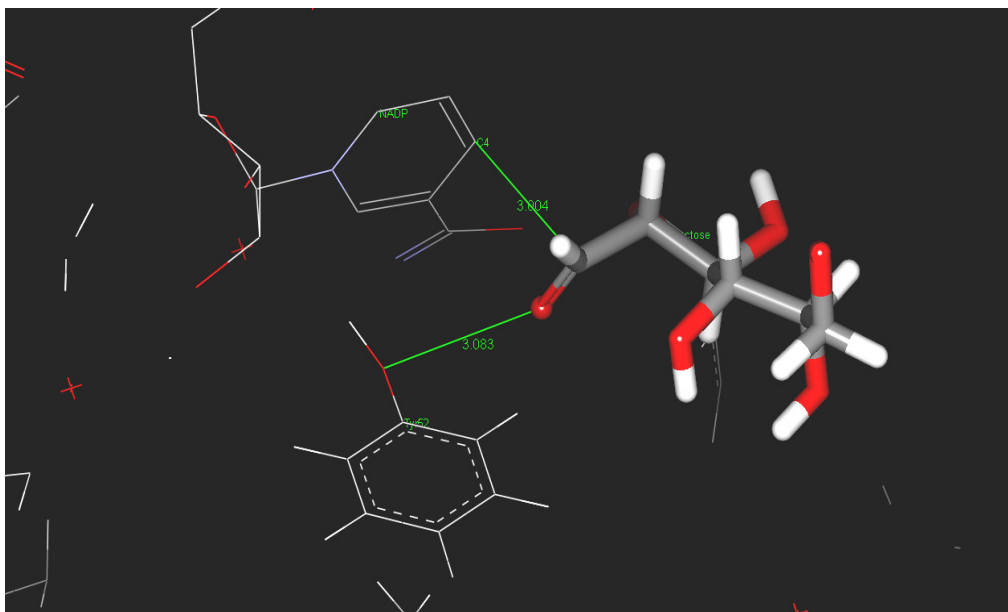
**Figure 4.8 Two potential conformations of methyl glyoxal found in the database**

- 4) Comparison of the orientation of the hits (the orientation of the aldehyde group and the configuration of hydroxyl group at C2 position, forming two H-bond interactions with the Asn310 obtained from the final pharmacophore and the X-ray structure modeled {136} with the D-Xylose, demonstrates close congruence between the two. This further assures the sensibility of the pharmacophore and verifies that such mode of orientation is important for the substrate recognition purposes within the active site.
- 5) Although short chain aldose compounds like D,L-Glyceraldehyde and D,L-Erythrose which have been identified as true substrates {138} were among



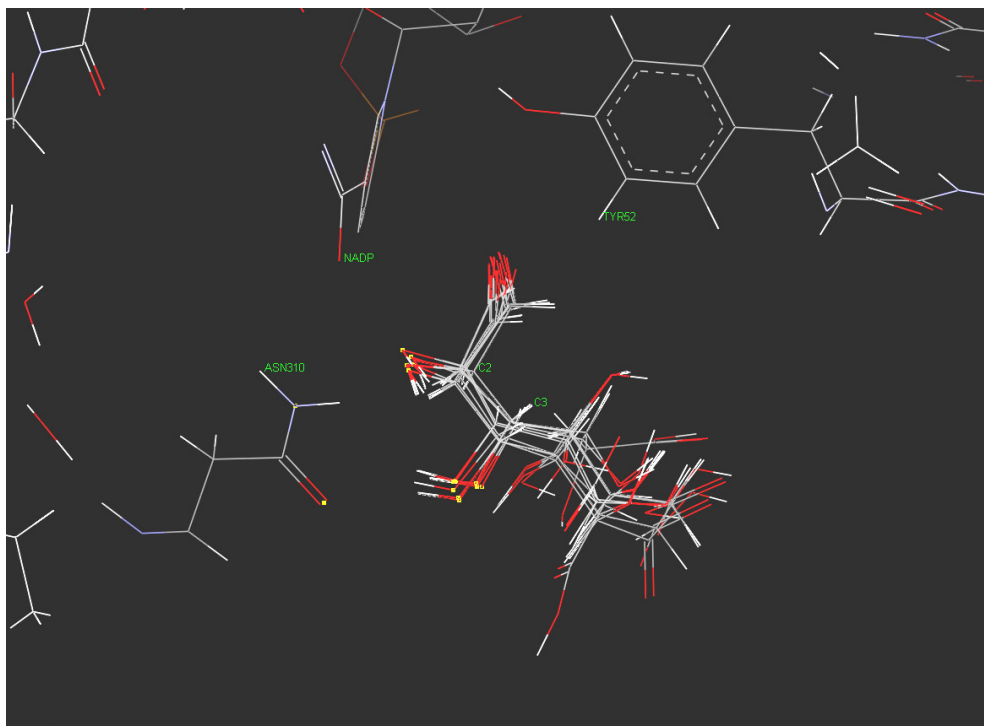
the hits in the initial pharmacophore models in which the number of constraints to be satisfied were less but as the number of constraints increased the short chain aldo compounds were unable to satisfy all the constraints and that's why they were not included in the hits from the final pharmacophore search. This clearly shows that having more constraints means that short chain aldehydes will be omitted from the result. One of the other reasons that some of the potential substrates have not been identified by the search as potential hits despite present in the database is because of the mode of sampling of different conformations of the same compound in terms of conformational energy of the individual conformation e.g the D-aldohexose has 100 conformations in the database and the range of the conformational energy of these conformations is 0-5.73 Kcal/mole. Similarly in case of D-galactose there are 100 possible conformations in the database and their energy range is 0-11.04Kcal/mol, therefore during the sampling of compounds with various conformations there is a probability that some of the compounds are compromised with respect to other compounds based on conformational energy sampling and thus losing them as a hit during search and resulting in missing some potential hits.

- 6) By visualizing hits from the final pharmacophore (Figure 4.9) it is observed that most of the hits have their aldehydic functional group oriented in such a manner, which brings both the carbonyl oxygen and carbonyl carbon at a close H-bond distance from the respective protein atoms. Further the carbon of the carbonyl carbon is able to get the hydride transfer from the Co-factor and thus making it plausible for the carbonyl oxygen to get the proton transfer from Tyr52 of the enzyme (Figure 4.8).



**Figure 4.9** Orientation of the D-galactose towards the co-factor NADP and Tyr52 in active site from pharmacophore searching model (image created by using DSV, certain residues are removed for clarity purposes)

- 7) A part from the alignment of the key aldehydic functional group there seems to be a specific trend among the hits for the hydroxyl group at C2 and C3 positions. For instance all the hydroxyl groups in D-galacturonate have exactly the same configuration as D-galactose (C2 up and C3 down for hydroxyl group), similarly in case of D-Ribose the hydroxyl groups at C2 and C3 are in agreement with D-galactose. Interestingly there is a visible preference of hydroxyl group in up configuration at C2 position in majority of the hits by making two H-bond interactions with the Asn310 of the enzyme (Figure 4.9). This has previously been reported in the modeled structure of XR with xylose {136}.



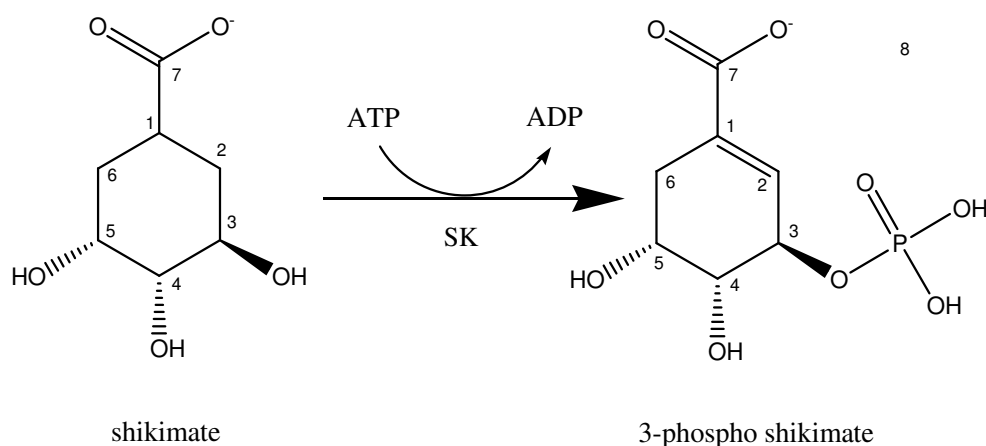
**Figure 4.10 Configuration of Hydroxyl groups on C2 and C3 of the majority of hits in the active site of XR(H-bond interactions are highlighted through yellow dots between hydroxyls of C2,C3 and N and O of Asn310, certain amino acid residues are removed for clarity,image created by using DSV)**

- 8) The hits obtained including certain aldo-pentoses such as D-Adipose and aldo-hexoses such as D-Allose, D-Fucose and D-Quinovose showing the same alignment as D-galactose and can be the potential substrate of the enzyme. This can be confirmed through the establish enzyme assay. From the work by neuhauser *et al* {138} which describes the relative activity of certain substrates e.g; that XR prefers glyceraldehyde, D-erythrose and even some aliphatic and aromatic aldehydes on pentose sugars such as D-xylose and L-Arabinose. The higher activity of the enzyme against 3-C and 4-C sugars is further supported by the fact that as there are more chances of hydration of 3-C and 4-C open chain sugars in comparison to the pentoses and hexoses (which tend to cyclise and have lesser chances to get hydrated, and thus the relative concentration of the open form is reduced). The lower activity of the enzyme against pentoses and hexoses can also be due to the absence of any mechanism by which the ring form of the compound goes to the active site and then converted to open chain form and further converted to product. This further shows that why the enzyme may have slow activity against these compounds. The higher

activity for small chain sugars could be due to the provision of higher number of H-bond interactions and thereby increasing their stability and recognition in the active site of the enzyme. From the specific orientation of majority of hits in the active site obtained through pharmacophore searching, it may not be wrong to suggest that the binding mode and activity of the substrates towards the enzyme active site is driven by the concentration of free aldehyde-form (open chain form) along with the activity of the chemicals (chemical features of the compound: H-bond donor & acceptor interactions).

## 5. Shikimate kinase form *Mycobacterium tuberculosis*

Shikimate kinase (SK; EC 2.7.1.71) is the fifth enzyme in the chorismate and shikimate biosynthetic pathway and catalyses the specific phosphorylation of the 3-hydroxyl group of shikimate to 3-phospho shikimate (Figure 5.1) in the presence of ATP as a co-substrate {96, 142}. In shikimate pathway, initially erythrose-4-phosphate is converted to chorismic acid in seven steps. Further chorismic acid is subsequently utilized for the biosynthesis of aromatic compounds such as p-amino benzoic acid, folate, ubiquinone and aromatic amino acids i.e.; tryptophan, phenyl alanine and tyrosine {143}. The shikimate pathway is of crucial significance in the metabolism of algae, fungi, bacteria and higher plants, but it is absent in mammals. Among humans, infectious diseases are the leading cause of deaths in the world and it has been learned that 90% of these deaths are connected with microorganisms. Among them *Mycobacterium tuberculosis* has the highest death ratio in comparison to other infectious agents {144} and this makes *M. Tuberculosis* SK (SK) a promising target for the development of non toxic novel anti-*M. Tuberculosis* drugs.



**Figure 5.1 Conversion of shikimate to 3-phosphoshikimate by SK with ATP as cosubstrate**

High resolution X-ray structures of SK with ADP and shikimate bound (PDB code:1U8A) {78} and one without shikimate (PDB:1LY4) {143} have been determined. The comparison of the two structures shows that the structure undergoes induced fit movement on substrate binding and conformational

changes occur particularly in the binding and lid domain of the structure after substrate binding.

## **5.1 Pharmacophore searching for SK**

The SK was chosen for pharmacophore searching method based on the following points.

1. Availability of high resolution structures
2. Availability of both ligand bound and unbound structures
3. Flexibility/modification in the structure especially in the active site
4. Single substrate specificity

So we were interested in seeing whether the pharmacophore searching method could identify the substrate based on interactions within the active site from the substrate bound structure and ultimately the substrate free case.

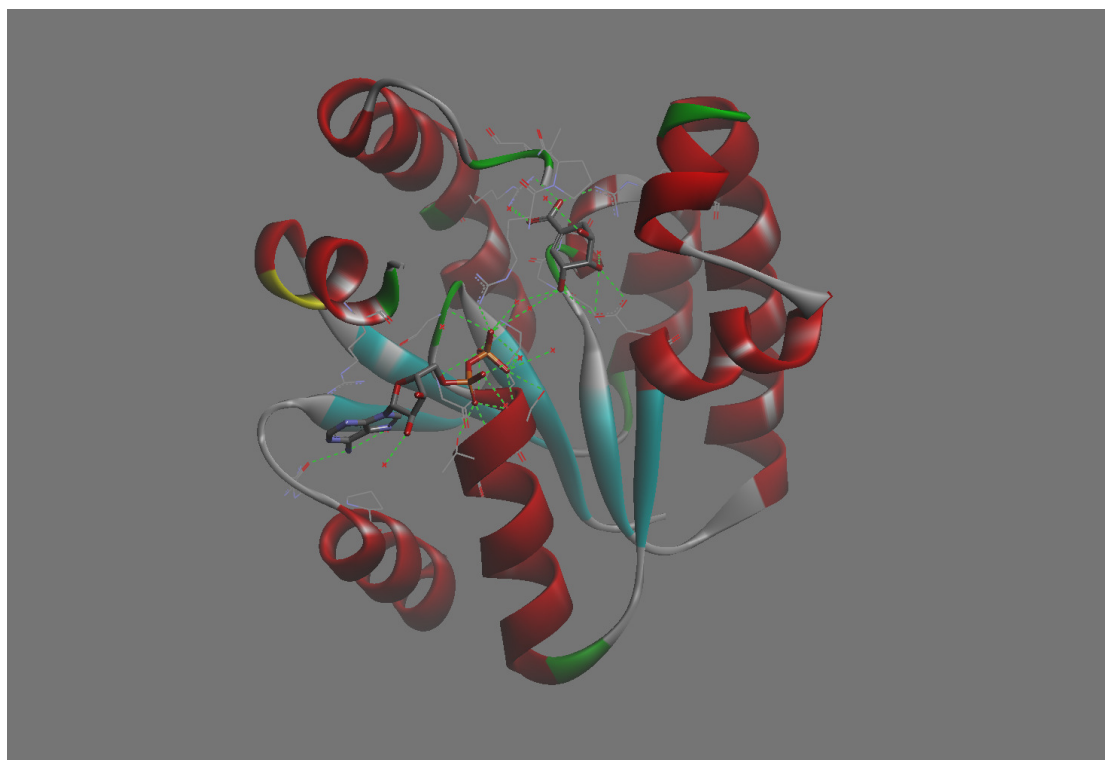
## **5.2 Aims and objectives**

The aims of this test case included the following points

1. To find out the optimum number of interactions needed to be used to reduce the number of hits to a manageable level
2. To compare the pharmacophore interactions and hits obtained among the substrate bound and unbound structures
3. To differentiate the vital interactions in the active site in the absence and presence of substrate
4. To identify parameters and interactions that need to be optimized to retain shikimate among the hits

### 5.3 Generation of Pharmacophore

Different operations were carried out for the generation and optimization of pharmacophore by using Accelrys Discovery Studio Visualizer® and Accelrys Catalyst® for database search. Naturalism database was used through out the search. The pharmacophore models were generated by using the DSV, Vector method as described in section 3.5.2. The H-bond donor and acceptor vectors were interconverted at various stages during the generation of pharmacophore models by editing the details of the .chm file in a text editor. In brief different approaches were used to explore the outcomes in the form of hits. These approaches included introduction of hydrogen bond donor and acceptor vectors, decrease and increase in the size of uncertainty location spheres around the head and tail of the hydrogen bond vectors and changing the radius around the active site for the introduction of exclusion sphere. Further details of subsequent steps and their outcomes are given in table 5.1. Figure 5.2 shows the solid ribbon diagram of SK (PDB: 1U8A) along with the co-factor ADP and product DHS forming important H-bonds with the active site residues.



**Figure 5.2** Solid ribbon diagram for SK along with ADP and DHS in the active site, forming Hydrogen bonds with the active site residues of the enzyme.

## 5.4 Pharmacophore generation via holo-enzyme

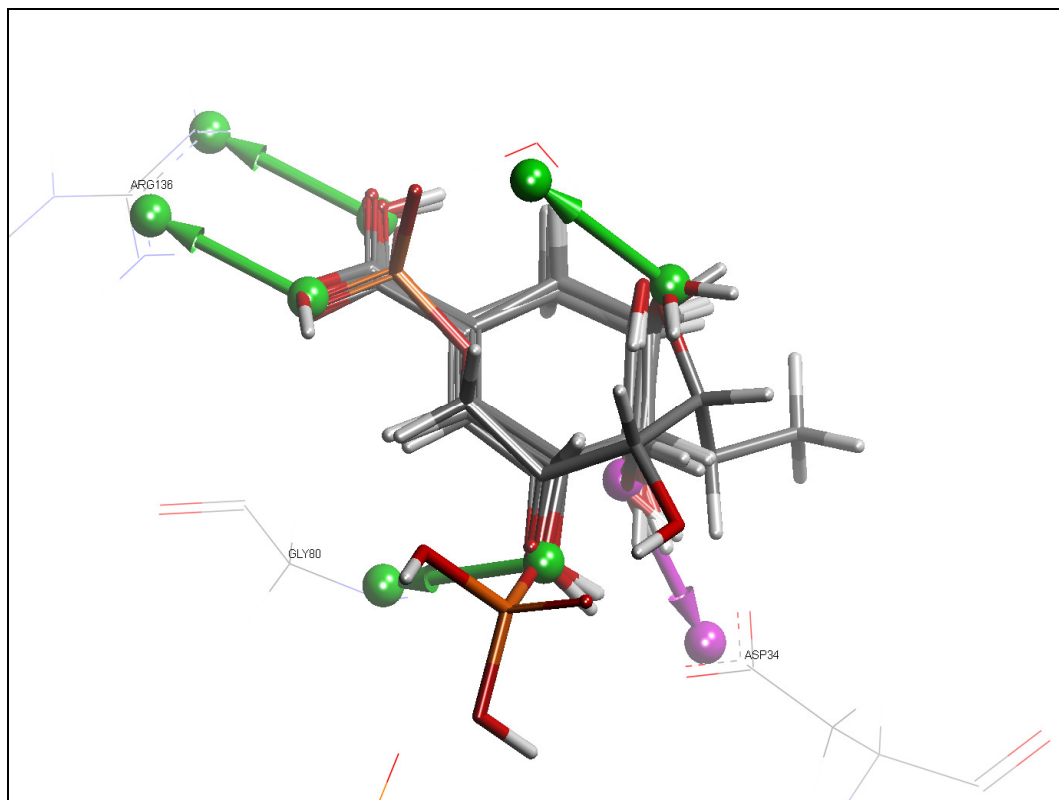
In the initial stages the holo-form of the enzyme structure (PDB code: 1U8A) was used for the generation of pharmacophore. All the pharmacophores produced through this structure were totally from the ligand (substrate) perspective. H-bond donor and H-bond acceptor vectors were introduced on the basis of possible interactions between the ligand and protein. The resultant pharmacophores were then subjected to database search. The step wise details for each pharmacophore are given below:

### 5.4.1 *Pharma Holo 1*

The first pharmacophore was generated by defining all the possible interaction in terms of the H-bond interactions provided by the shikimate. In total 4 H-bond acceptor vectors were introduced, 2 from the carboxylate group of C7 of shikimate towards the guanidinium group of ARG136 and 2 from hydroxyl groups of C3 and C5 of shikimate towards the amino end of GLY80 and water molecule respectively. An additional single H-bond donor vector was introduced from hydroxyl group of C4 of shikimate towards the carboxylate of ASP34.

Most of the hits obtained as a result of this pharmacophore were correct. The hits included both the reactants (shikimate) and products (shikimate phosphate) of the actual shikimate kinase reaction (Figure 5.2). 5 of the hits were phoso-shikimate compounds (which is the product in the actual reaction). Some of the carbohydrates were also included in the hits possibly because of their flexibility to fit in and satisfy the pharmacophore. Among hits, some of the carbohydrates had their phosphate group oriented toward the carboxylate of the shikimate. This highlighted the need to define a chemical group such as a carboxylate at this position in the pharmacophore to avoid the false hits.

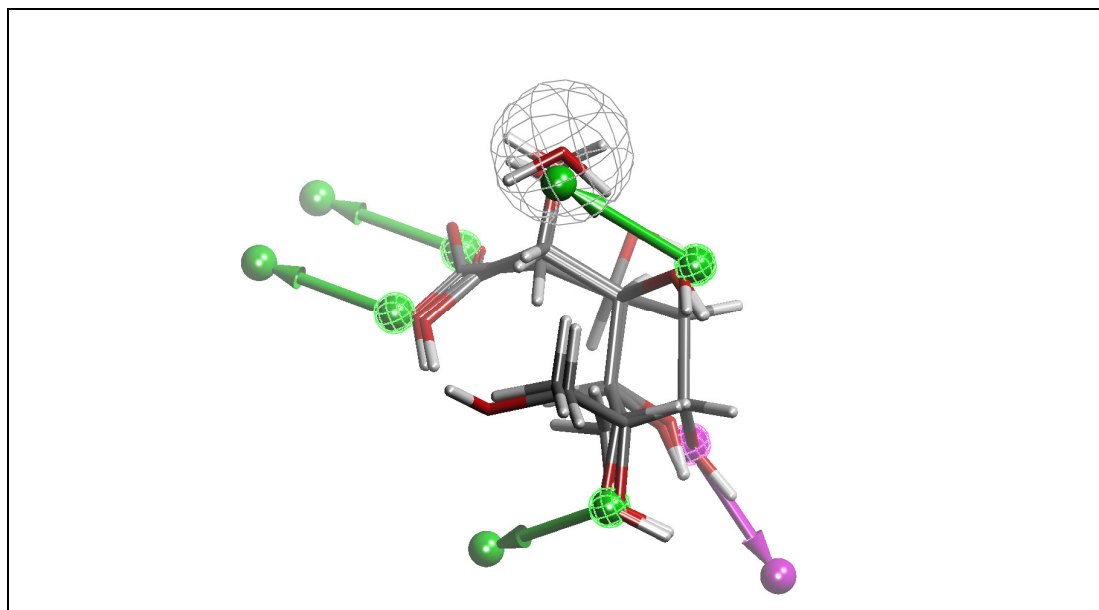




**Figure 5.3** The hits obtained include both the shikimate and shikimate phosphate and their derivatives in the same orientation as described in the X-ray structure of holo-form of SK (Green Vector = Hydrogen bond Acceptor, Purple Vector = Hydrogen bond donor)

### 5.4.2 Pharma Holo 2

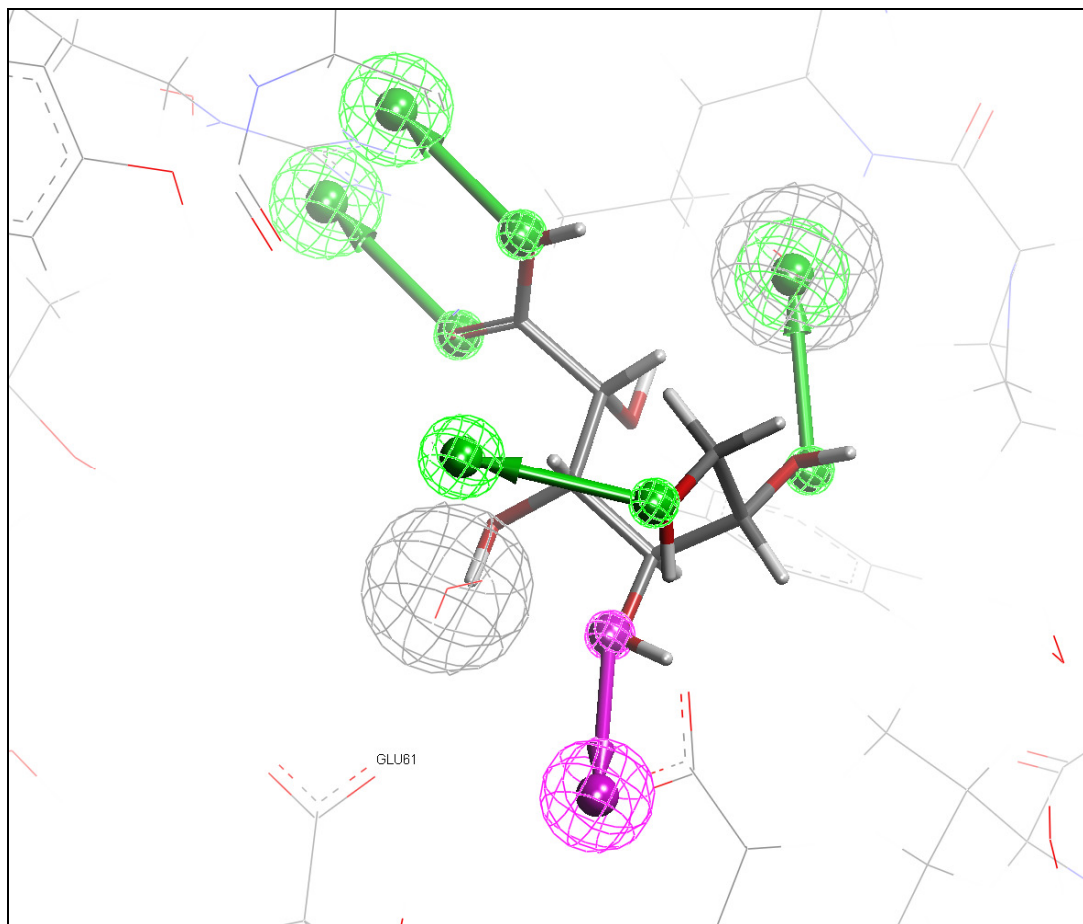
In order to avoid the steric clash of the hits with the water molecule, an exclusion sphere was introduced around the water molecule providing H-bond acceptor interaction from the hydroxyl group of C5 of shikimate (Figure 5.2). The resultant search gave hits which did not include carbohydrates like D-fuconic acid, D-galactonate, D-galactonic acid. The mentioned compounds were not selected possibly due to the clash between the hydroxyl group of these compounds with the exclusion sphere around the water molecule (Figure 5.3).



**Figure 5.4** the introduction of exclusion sphere around the water molecule leads to lose some hits due to possible steric clash (the compounds shown were not selected among the hits).

### **5.4.3 Pharma Holo 3**

In order to further avoid the false hits an additional constraint was added by introducing an exclusion sphere around a water molecule lying between shikimate and GLU61. This resulted in reducing the number of hits from 14 to 10. It appeared that the compounds like L-idonic acid and L-idonate were not selected because their hydroxyl groups at C3 position was in direct clash with the exclusion sphere of the water molecule (Figure 5.4).



**Figure 5.5** The hydroxyl group of the hit (L-idonic acid) in direct clash with the exclusion sphere of the water molecule close to GLU61, the image position has been re-oriented for viewing purposes (the image has been reoriented for viewing purposes).

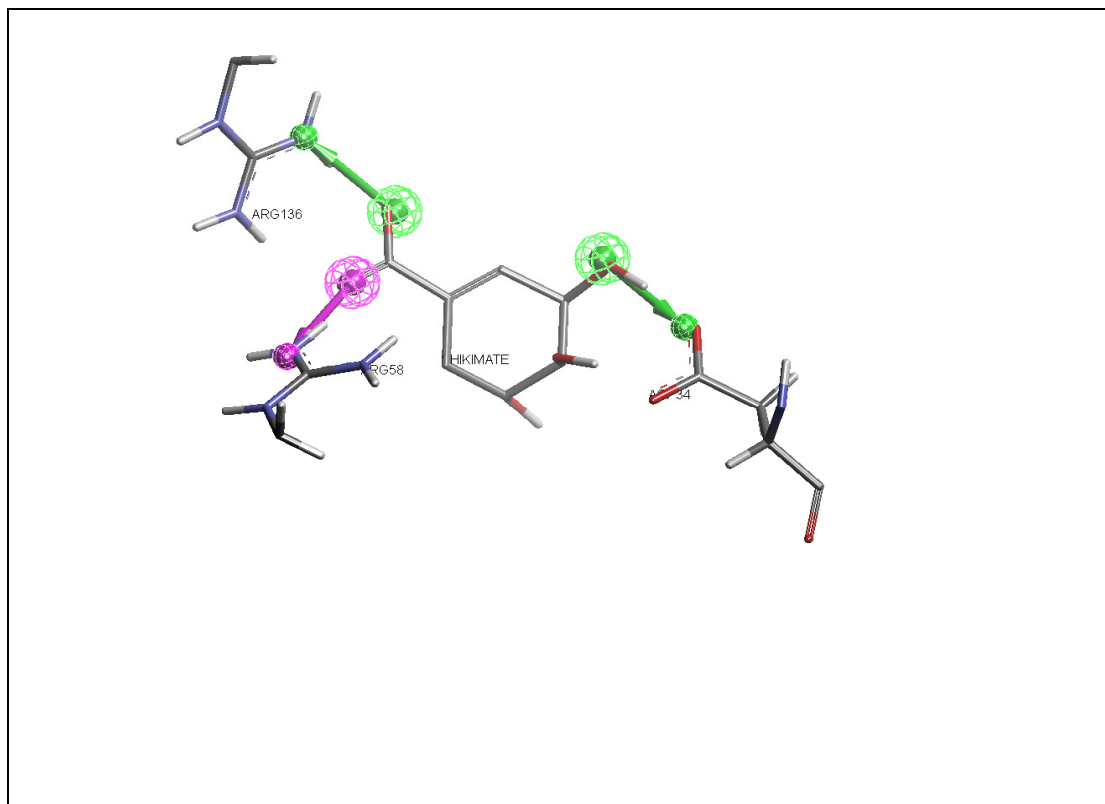
#### **5.4.4 Pharma Holo 4-5**

To explore the critical interactions and their effects in the form of different hits the H-bond acceptor vector pointing from C3 hydroxyl group of shikimate towards GLY80 was changed to H-bond donor vector. As a result the donor vector now pointed towards the carboxylate group of ASP34. The resulting hits were reduced to 4, including shikimic acid, shikimate, D-erythritol 4-phosphate and D-arabinose 5-phosphate. The phosphate groups of the two compounds were oriented towards the carboxylate group of shikimate. However, the product shikimate phosphate was lost. The results from the pharmacophore indicated that the constraints were too strict in some ways. Further the size of uncertainty location spheres for the atomic positions was far too small to be practically useful which needed to be relaxed. However, the pharmacophore exhibited the type of H-bond interaction vectors required to get the minimum number of hits

along with the true substrate. The H-bond donor interaction from the hydroxyl group of C3 of shikimate was not very fruitful therefore in Pharma Holo 5 the pharmacophore model was reverted back to Pharma Holo 3 model. The angle and direction of the H-bond vectors were optimized to point exactly toward the target atom.

#### **5.4.5 Pharma Holo 6**

In order to explore the critical interactions in the active site of the shikimate kinase, H-bond interactions were removed from the hydroxyl groups of C3 and C4 of the shikimate. Further the H-bond acceptor interaction from the carboxylate group of C1 of shikimate was changed to a H-bond donor, thus pointing towards guanidinium group of ARG58 instead of ARG136. The orientation of H-bond acceptor vector originating from the hydroxyl groups of C5 of shikimate was optimized, so that it pointed towards carboxylate group of ASP34 instead of a water molecule. The radius of the uncertainty locations spheres around the blobs of tails and heads of the H-bond vectors was changed to 0.6Å and 0.3Å respectively (Figure 5.5). The numbers of hits obtained were suddenly increased to 290, showing that a lower number of constraints leads to a higher number of hits from the database search.



**Figure 5.6** Three H-bond interactions, one donor and one acceptor from the shikimate carboxylic group toward Arg136 and Arg58, and one acceptor from C4 hydroxyl group toward Asp34 (the image has been reoriented for viewing purposes).

#### **5.4.6 Pharma Holo 7**

Further optimization of the pharmacophore was carried out by changing H-bond acceptor vector to H-bond donor vector pointing from C5 hydroxyl group of shikimate towards the carboxylate group of ASP34. Along with it the H-bond donor vector was converted to H-bond acceptor vector pointing from carboxylate group of shikimate towards the guanidinium group of ARG58. The pharmacophore hits were further increased to 437, demonstrating that fewer interactions to act as constraints result in more hits which can satisfy them.

#### **5.4.7 Pharma Holo 8**

In order to reduce the number of hits without losing the true substrate, the pharmacophore model was reverted back to Pharma Holo 6 model. The pharmacophore was edited by the introduction of H-bond acceptor vector from the C4 hydroxyl group of shikimate towards the carboxyl end of ASP34. Radius of uncertainty location sphere around the vector tail and head were set to 0.3Å and

0.6Å respectively. Though the numbers of hits were reasonably reduced to 87 but the hits did not include shikimate, showing that the inclusion of exclusion sphere around the water molecule and/or the inclusion of new H-bond acceptor vector in the pharmacophore are not suitable for selection of shikimate as a hit.

#### 5.4.8 Pharma Holo 9-10

The whole model was kept the same as Pharma Holo 8 model, except the radius of uncertainty location spheres around the blobs of tails and heads of H-bond vectors (both acceptors and donors) was changed from 0.6Å to 0.3Å (Figure 5.6A). The pharmacophore resulted in 4 hits without shikimate, showing that not one but two H-bond acceptor vectors pointing towards the guanidinium group of Arg136 are vital for recognition of true substrate (as shown in Pharma Holo 3). Later on in Pharma Holo 10 the radius of uncertainty location spheres around the blobs of heads and tails of H-bond vectors (both acceptors and donors) was changed back to 0.6Å (Figure 5.6B).

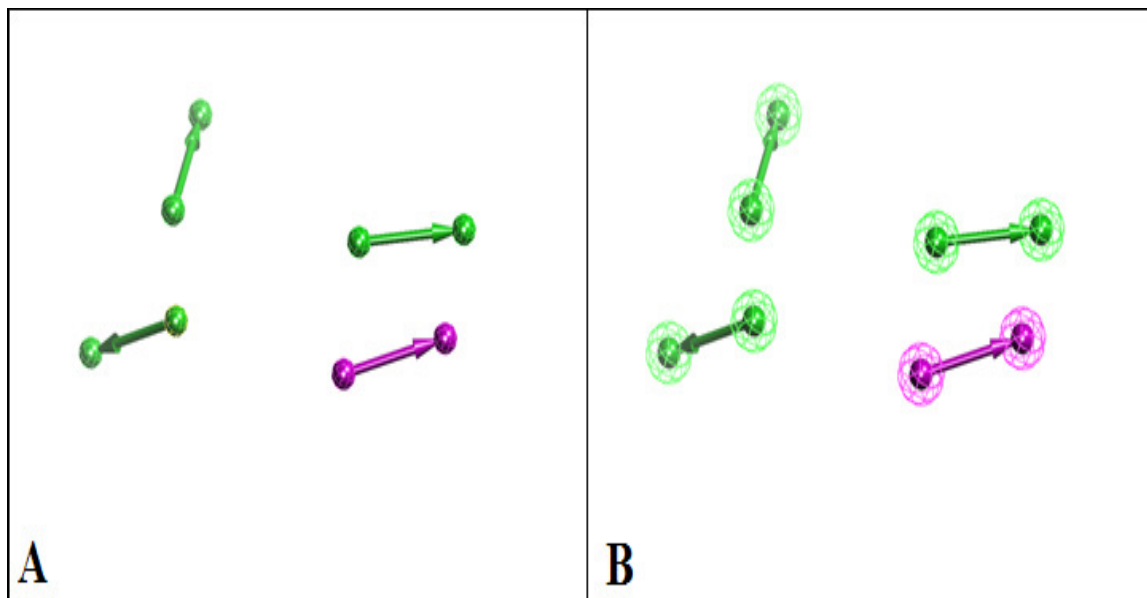


Figure 5.6 The radius of the uncertainty location spheres both around the head and tail of the H-bond vectors is 0.3Å in A and 0.6Å in B

### **5.4.9 Pharma Holo 11-14**

Further optimization of the pharmacophore was carried out by changing the radius of the uncertainty location sphere for H bond vectors (both acceptors and donors). The radius around the blobs head pointing towards the amino acid end were fixed to 0.6Å and for the blob tail originating from the ligand end were changed from 0.6Å to 0.3Å (Pharma Holo 11). Later on in Pharma Holo 12 the radius was further changed from 0.6Å to 0.8Å for the blob head and in Pharma Holo 13 the radius around the blob tail of H-bond vectors was changed from 0.3Å to 0.35Å. The relaxation in the radius caused significant increase in the number of hits as shown in table 5.1. The pharmacophore models from Pharma Holo 7 to Pharma Holo 13 had a H-bond acceptor interaction pointing from the carboxylate of shikimate towards the guanidinium group of Arg58 and none of these pharmacophores were able to pull out shikimate as a hit from the database, but as soon as in Pharma Holo 14 model this H-bond acceptor interaction is directed toward the guanidinium group of Arg136 instead of Arg58, the pharmacophore readily selects shikimate as a hit. This showed that interactions provided by the Arg136 in the pharmacophore search model are vital for the recognition of true substrate among the hits via database search.

### **5.4.10 Pharma Holo 15-16**

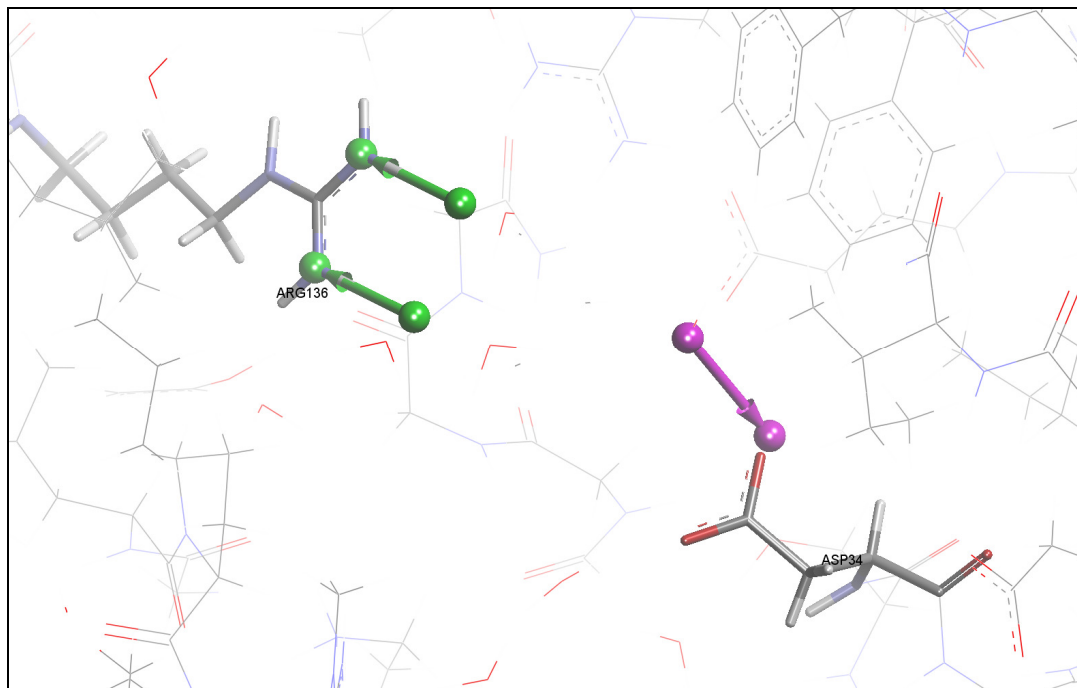
Keeping in all the constraints same as in Pharma Holo 14 model, only the radius of uncertainty location spheres around the blob head of both H-bond acceptor and donor vectors was changed from 0.8Å to 0.6Å. This change still kept the shikimate among the hits, thus exhibiting the validity of the pharmacophore. To check the requirement of the optimum radius for uncertainty location spheres around the blob head of H-bond vectors (both acceptors and donors) the radius was further reduced to 0.4Å in Pharma Holo 16. This resulted in losing shikimate as a hit, demonstrating that higher stringency in constraints in terms of smaller radius of the uncertainty location sphere would lose the true substrate through the database search.

## 5.5 Apo-enzyme for pharmacophore generation

By going through various optimization stages of generating pharmacophore by using the holo-enzyme structure, it became obvious that certain optimized constraints are crucial and need to be in the pharmacophore model for selection of true substrate among the hits. Keeping in view these constraints the pharmacophore was generated by using the apo-structure of the enzyme (PDB code: 1L4Y). The pharmacophore models were generated by using the DSV, Vector method as described in section 3.5.2. During the course of optimization various pharmacophore models were generated by using the apo-structure of the enzyme, further details are given below.

### 5.5.1 Pharma Apo 1

Keeping in view the vital interactions offered by certain amino acids in the active site, in the initial pharmacophore 2-H bond acceptor vectors were introduced towards the guanidinium group of ARG136 and 1 H-bond donor vector pointing from a water molecule towards the carboxylate group of ASP34 (Figure 5.7). The hits ( $\geq 300$ ) included shikimate showing that the H-bond interactions from Arg136 and Asp34 are vital for substrate recognition.



**Figure 5.7** The generation of pharmacophore by using the apo-structure (PDB code: 1L4Y) Arg136 and Asp34 are shown as stick models with the corresponding H-bond acceptor (green arrows) and donor (purple arrows) vectors respectively.

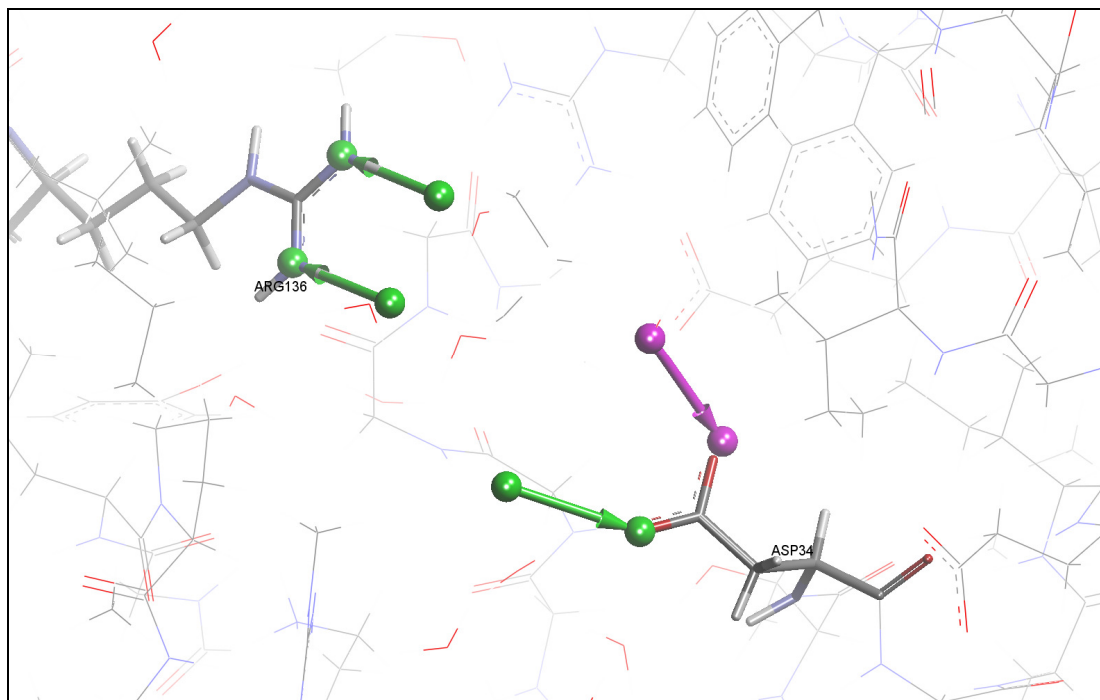


### 5.5.2 Pharma Apo 2

Further to the Pharma Apo 1 model, the radius of the uncertainty location spheres around the blob of tail of the vectors originating from the potential ligand side were changed from 0.6Å to 0.35Å. This stringency in the constraint still held the true substrate (shikimate) among the hits ( $\geq 300$ ).

### 5.5.3 Pharma Apo 3

To further probe the important interaction provided by the active site of the enzyme, an additional H-bond acceptor vector was added from the carboxylate end of Asp34 (Figure 5.8). This additional interaction fitted well in the pharmacophore and reduced the number of hits (123) without losing the true substrate (shikimate).



**Figure 5.8** Pharmacophore generated by using the apo-structure (PDB code: 1L4Y) with the additional H-bond acceptor interaction (green arrow) originating from the carboxylic group of Asp34.

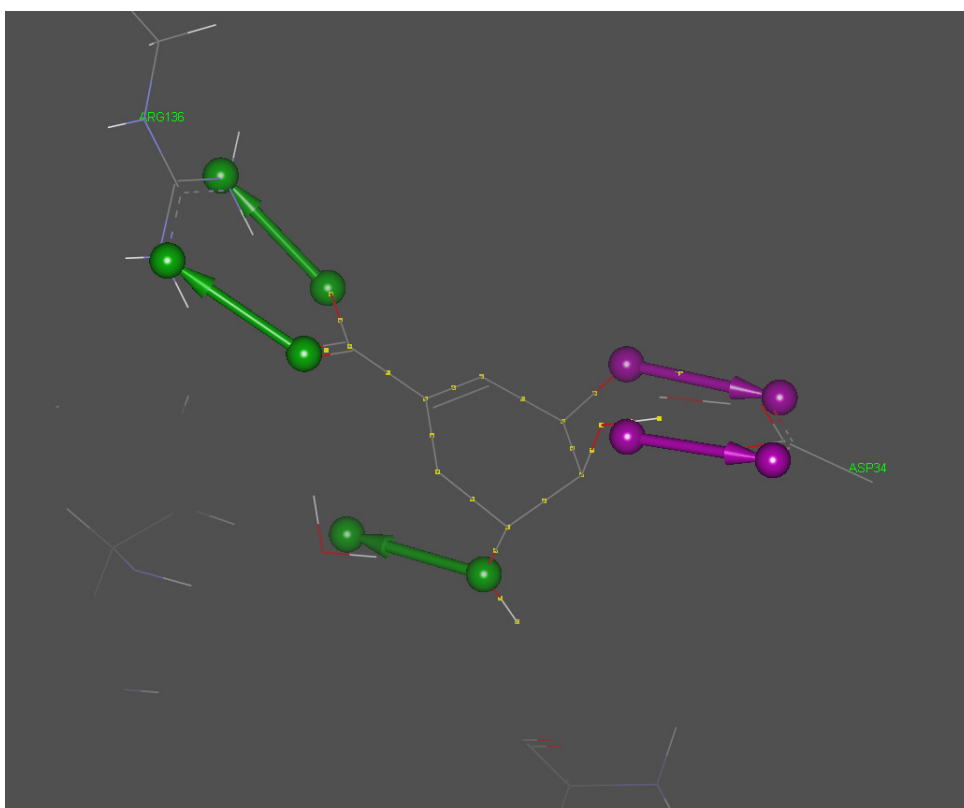
**Table 5.1 Total number of constraints used in each pharmacophore model, along with quantity of hits in each case and time taken for database search (table shows the presence of the true substarte (shikimate) among the hits in each pharmacophore model search).**

Pharmacophore model	Total number of constraints	Number of hits obtained	Database search time (minutes)	Shikimate among the hits (Yes/No)
Pharma Holo1	6	18	35	Yes
Pharma Holo 2	7	14	25	Yes
Pharma Holo 3	8	10	50	Yes
Pharma Holo 4	8	4	20	Yes
Pharma Holo 5	6	18	35	Yes
Pharma Holo 6	4	290	20	Yes
Pharma Holo 7	4	437	10	Yes
Pharma Holo 8	6	83	18	No
Pharma Holo 9	6	4	13	No
Pharma Holo 10	6	209	47	No
Pharma Holo 11	6	38	20	No
Pharma Holo 12	6	53	28	No
Pharma Holo 13	6	89	36	No
Pharma Holo 14	6	110	28	Yes
Pharma Holo 15	6	75	19	Yes
Pharma Holo 16	6	25	20	No
Pharma Apo 1	4	≥300	55	Yes
Pharma Apo 2	4	≥300	10	Yes
Pharma Apo 3	5	123	45	Yes

## 5.6 Conclusions

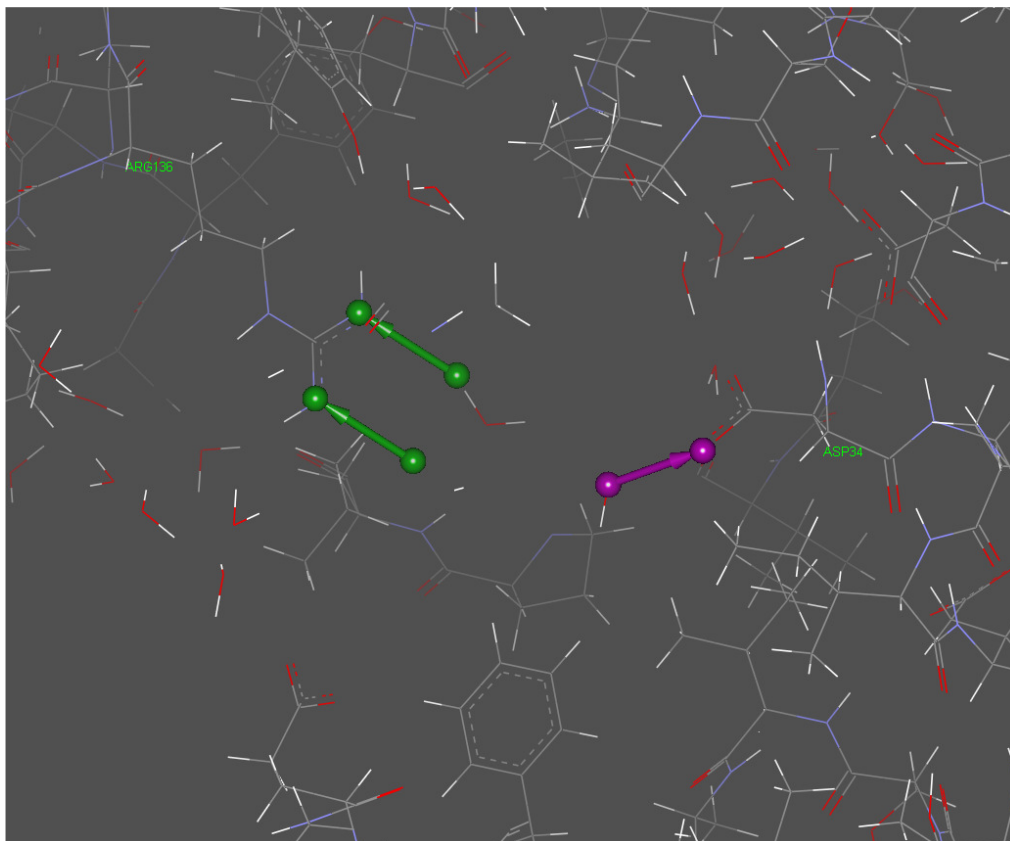
The course of optimization of pharmacophores for SK structure both with and without the substrate bound form demonstrates that certain interactions are vital for the selection of shikimate to come out as a hit. Based on the results obtained through optimization of various pharmacophore the following conclusions can be drawn:

- 1) In case of the pharmacophore generation, when using the substrate bound structure the Pharma Holo 4 demonstrates the minimum required number of constraints for the limitation of hits to a minimum number with shikimate present as a hit. It appears that the vital interactions include the two H-bond donor vectors pointing towards the ASP34 and 3 H-bond acceptor vectors, 2 pointing towards the guanidinium group of ARG136 and one towards the water molecule (Figure 5.9).



**Figure 5.9** Crucial interactions required for the selection of shikimate as a hit from the database, when using the substrate bound structure for the generation of pharmacophore, two H-bond donor interactions(in magenta) towards the carboxylate of ASP34, and three H-bond acceptor interactions(in green), two towards the guanidinium group of ARG136 and one pointing towards the water molecule, shikimate structure is highlighted with yellow dots(exclusion and uncertainty spheres are removed for clarity purposes)

- 2) In case of using the apo structure of the enzyme, the crucial interactions required for the selection of shikimate among the hits include the H-bond donor interaction pointing towards the carboxylate group of the ASP34 and 2 H-bond acceptor interactions pointing towards the guanidinium group of ARG136. This can be noticed in the course of optimization of the pharmacophore models in Pharma Apo 1,2 and 3 and also depicted in figure 5.10



**Figure 5.10** Crucial interactions required for the selection of shikimate as a hit from the database, H-bond donor interaction(in magenta) towards the carboxylate of ASP34, and two H-bond acceptor interactions(in green) towards the guanidinium group of ARG136(exclusion and uncertainty spheres are removed for clarity purposes)

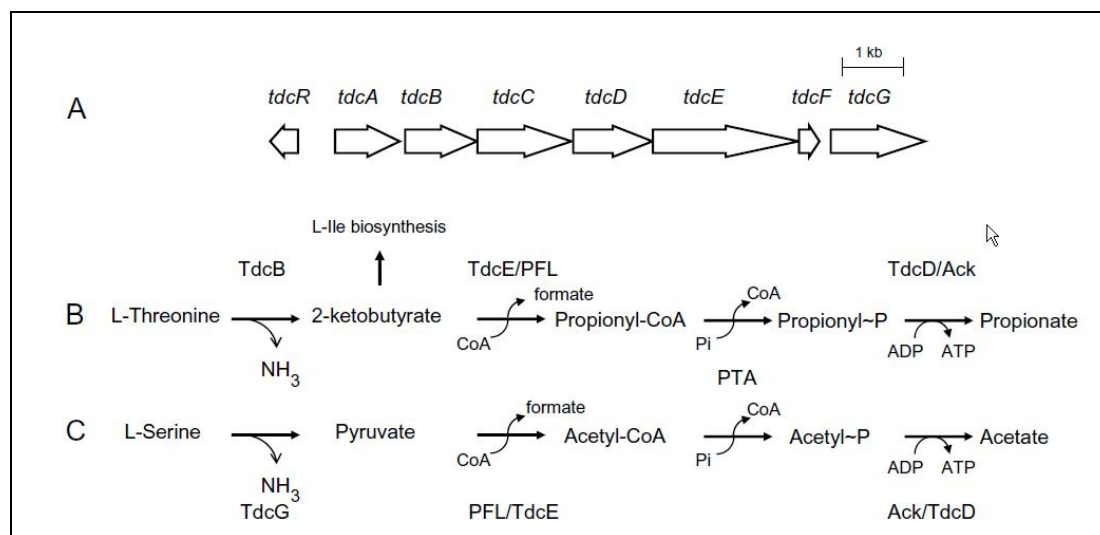
- 3) By comparing the two pharmacophores generated separately from the substrate bound structure and the apo form of the enzyme, it seems obvious that the ARG136 and ASP34 are crucial in providing the H-bond interactions for the recognition and stabilization of the shikimate within the active site and for the transfer of phosphate group to it.

- 4) It appears from the optimization of the pharmacophore searching that the size of the uncertainty location spheres around the H-bond head and tail is important in the selection of the true substrate as a hit. The optimization of pharmacophore during Pharma holo 15 & 16 indicates that the optimum radius of the uncertainty location spheres should be 0.6Å towards the blob head end and 0.3Å towards the blob tail end to avoid losing the true substrate (shikimate).
- 5) It appeared that the problem with shikimate kinase apo structure was that, key main chain interactions are not evident without knowledge of the ligand bound structures and therefore restricting the number of hits was a problem. The inability at this stage in the project to define chemical groups also resulted in obvious false miss hits that polluted the solution results.

## 6. The *Escherichia coli* protein *TdcF*, Pdb code: 2UYN

### 6.1 Introduction

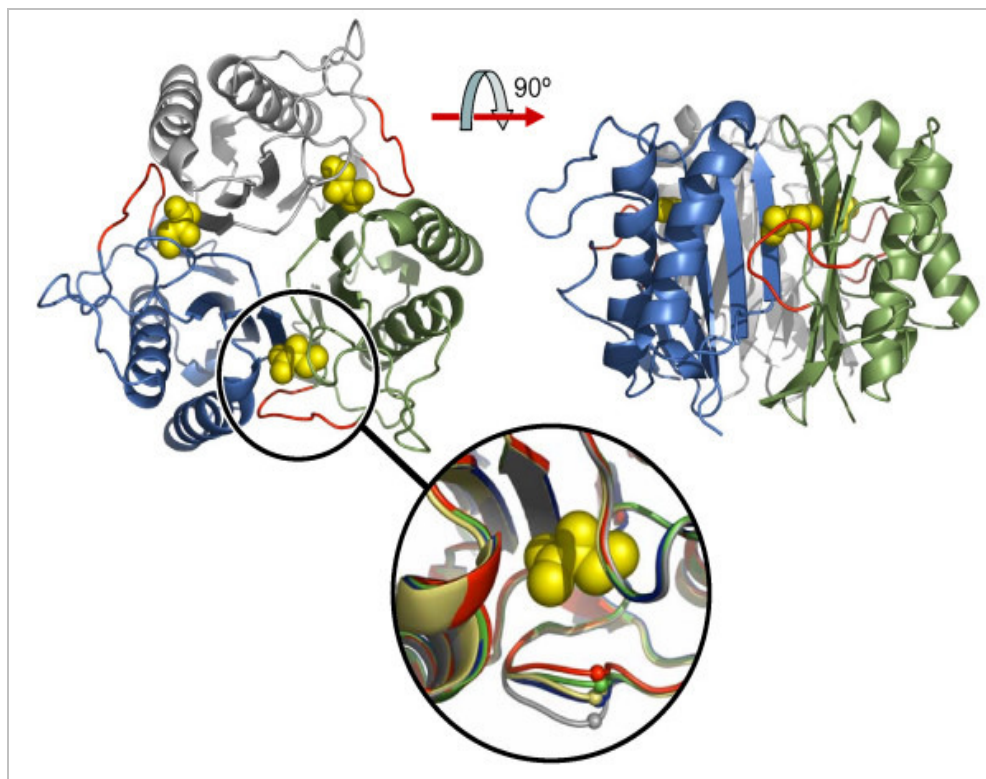
The *E. coli* protein *TdcF* is a member of the highly conserved YjgF/YER057c/UK114 family of proteins, found in euarchaea, higher plants and prokaryotes. The protein is widespread in nature but as yet has no clear biological function. The *tdc* operon shows the presence of *tdcF* gene among other genes like *tdcC* whose product is L-threonine/L-serine permease. On the basis of binding interactions of some of the analogues of *TdcF* with keto acids and the co-crystallized structure of *TdcF* with 2-ketobutyrate (Figure 6.1), its novel role in the metabolism of 2-ketobutyrate and in turn L-Ile biosynthesis has been suggested [145]. To date no binding data is available for *TdcF* to show weak or tight binding against different ligands.



**Figure 6.1 Representation of the sequence of genes on the *tdc* operon on the *E-coli* chromosome and the function of the corresponding gene products, along with the pathways for degradation of L-threonine and L-serine, (A) *tdc* operon, (B) L-threonine anaerobic degradation pathway, and (C) L-serine degradation pathway, Ack = acetate kinase; PFL= pyruvate formate-lyase; PTA = phosphotransacetylase. Image adapted from [145].**

The crystal structure of *TdcF* alone and in complex with a number of ligands has been solved by Lawson and co workers [145]. This confirms that the protein has a chorismate mutase like fold and exists as a homotrimer (Figure 6.2). There are large intersubunit clefts in which the ligands are bound within a well defined cavity. No catalytic activity has been ascribed to any member of this protein

family. It has been suggested on the basis of structural genomics and structural characteristics that 2-ketobutyrate is the best binder to the active site among other low molecular weight compounds {145}.



**Figure 6.2** Line ribbon diagram/model of *TdcF* protein, showing the binding site in zoom, image adapted from {145}.

All the ligands in the x-ray structure leave a large empty part in the cavity of the active site with most of it occupied by water molecules (Figure 6.3). Therefore the protein was chosen as a test candidate to find out the potential ligands by using pharmacophore searching. This might reveal other potential ligands which could occupy more fully the binding cleft, and which could answer certain questions like

- 1) If there is a possibility of binding of a larger ligand?
- 2) If a second potential ligand/cofactor can fit in?
- 3) If the binding site confers potential capability of protein-protein interactions?

All these queries make 2UYN a good test case and the availability of ligand binders means that the binding of potential ligands using different biophysical techniques can be evaluated. Finally the presence of different ligands structures

can significantly help to describe the potential ligand binding interactions for pharmacophore modeling.

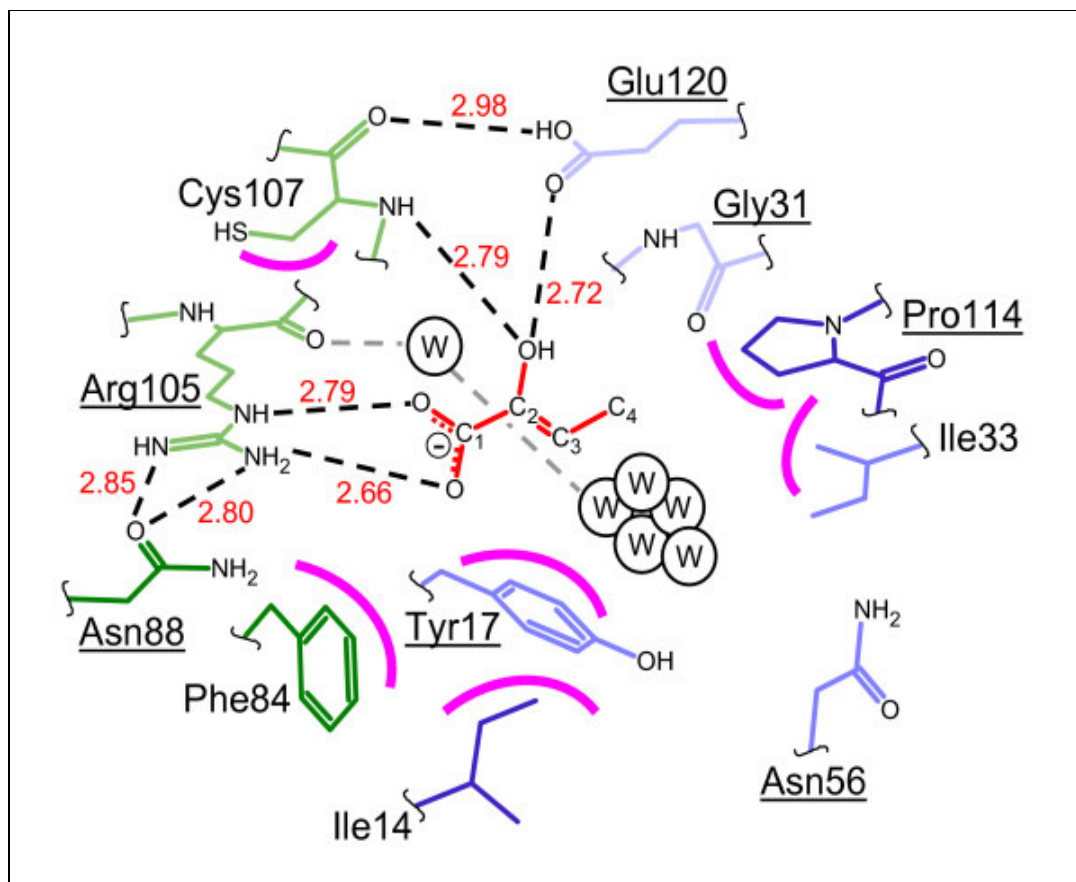


Figure 6.3 Binding site of the protein showing the interactions with 2-ketobutyrate. (A large part of the binding pocket is occupied by water molecules), W = Water molecules, image adapted from {145}.

## 6.2 Generation and Optimization of pharmacophore for *TdcF*

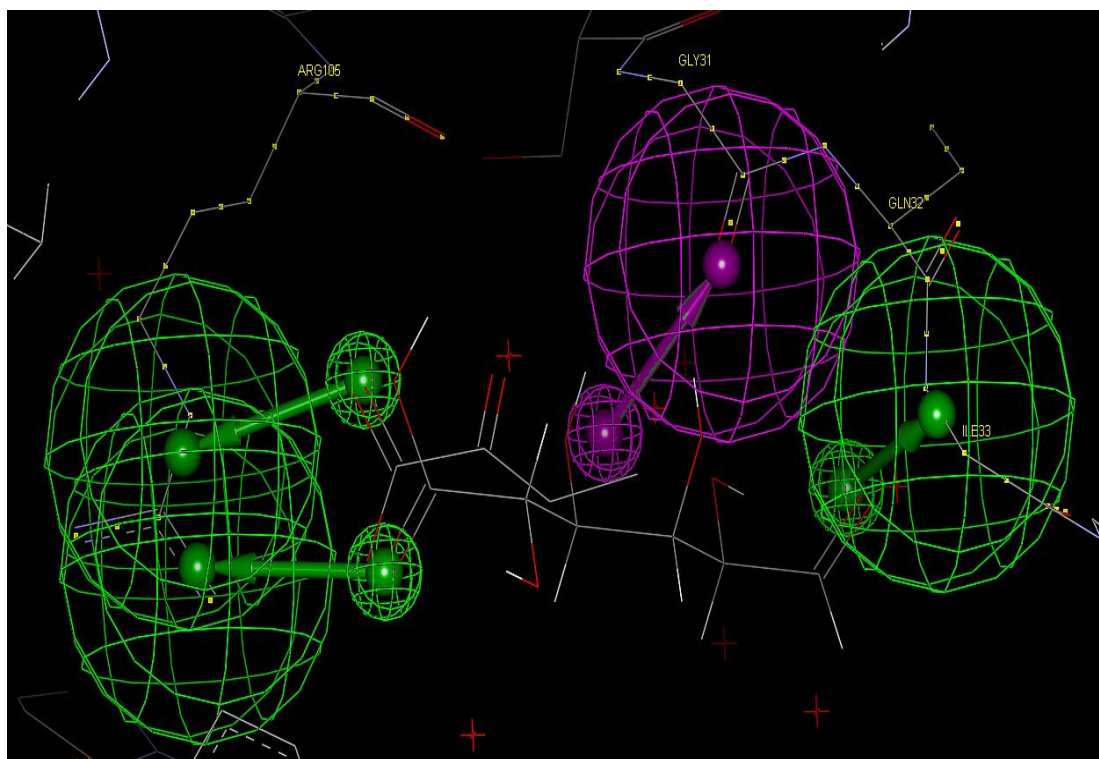
A number of different pharmacophore models were developed as we learned what would make an effective pharmacophore given the limitations of the methods available. The different models and steps taken are given below:

### 6.2.1 Pharma *TdcF* 1

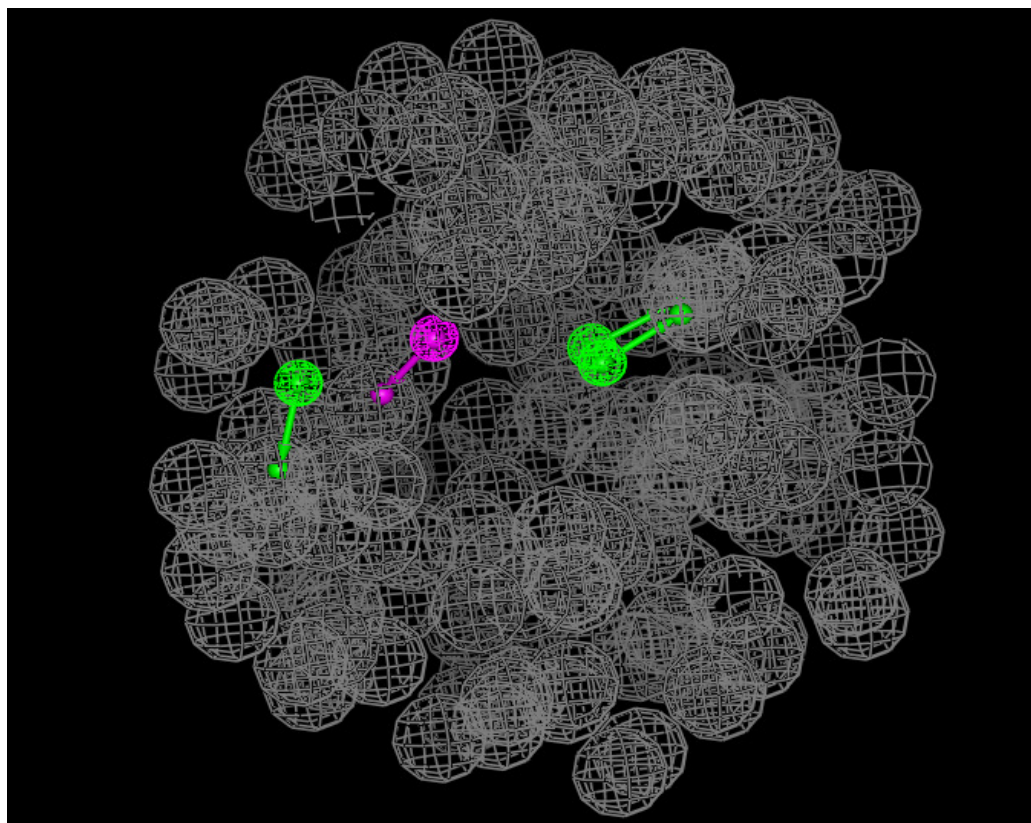
First pharmacophore was designed by using Cerius 2 and Weblab Viewer Pro programs as described in section 3.5.1. By using the *TdcF* X-ray structure water molecules and the ligand were removed from the active site. Hydrogens were generated for all the atoms within a radius of 14.0Å around the binding site by using the program Cerius2 {69}. Exclusion spheres were generated for all the non



hydrogen atoms and an interaction map was calculated. From this interaction map, individual interactions were chosen to represent hydrogen bond donors and acceptors which were consistent with the main interactions seen in the protein-ligand structure {145}. Initially three hydrogen bond acceptor vectors (green) and 1 hydrogen bond donor vector (purple) were introduced (Figure 6.4). Out of 3 hydrogen bond acceptor vectors, 2 were introduced from the basic guanidinium group of the Arg105 and 1 hydrogen bond acceptor vector was introduced from the peptide nitrogen of the GLN32 and ILE 33 peptide bond. The only H-bond donor vector was introduced from the carboxyl oxygen of the GLY31 (Figure 6.4). The exclusion spheres were introduced around 10Å radius of the binding site to prevent the direct space clash between the protein and potential hits. The size of uncertainty spheres around H-bond vectors was kept 2.2Å towards the protein end and 0.6Å towards the ligand end.



**Figure 6.4** Pharamcophore with 3 H-bond acceptors and 1 H-bond donor vector pointing towards the amino acid residues(yellow dotted structure) in the binding site with some of the hits obtained from the database search (exclusion spheres are removed for clarity)



**Figure 6.5** The initial pharmacophore generated, with hydrogen bond acceptors (green) and a hydrogen bond donor (magenta), exclusion spheres are shown as gray spheres.

The generated pharmacophore was searched against a multi-conformational library of metabolites (Naturalism database with 5492 small ligands) by using the Catalyst program (Accelrys). The search resulted in 178 potential ligands which satisfied the restraints and H-bond features as defined in the pharmacophore model. The ligands could be compared not only to the pharmacophore but also to the original crystal coordinates. This allowed further refinement and optimization of the pharmacophore positions using WebLab Viewer Pro (Accelrys). The numbers of hits obtained from the naturalism data base were 178 and search time was 123 minutes

### 6.2.2 Pharma *TdcF* 2

The hits obtained from first pharmacophore showed that most of the hits with their side chains were at H-bond distance from the carboxyl oxygen of the ARG105. Among the hits least H-bond interactions were seen for GLY31. Two changes were carried out, first the size of the uncertainty spheres around the H-

bond vectors towards the protein end was changed from 2.2Å to 1.2Å, and secondly the position of H-bond donor was changed from GLY31 to the carboxyl oxygen of the ARG105. The naturalism database search gave 65 hits in 27 minutes.

### 6.2.3 Pharma *TdcF* 3

Most of the hits obtained from the previous model were a bit short of H-bond distance and in some cases the atoms of the hits were not exactly pointing towards the tail of hydrogen bond vector (Figure 6.6). Therefore for the two hydrogen bond acceptors pointing towards ARG105, the x,y,z co-ordinates were optimized as described in section 3.5.1. The length of the H-bond acceptor vectors towards ARG105 was changed from 2.3Å to 3.0Å. The length of the H-bond acceptor pointing towards the ILE33 was changed from 2.4Å to 3.1Å. For H-bond donor vector the length was changed from 2.4Å to 2.8Å. 88 hits were obtained from the database search in 77 minutes.

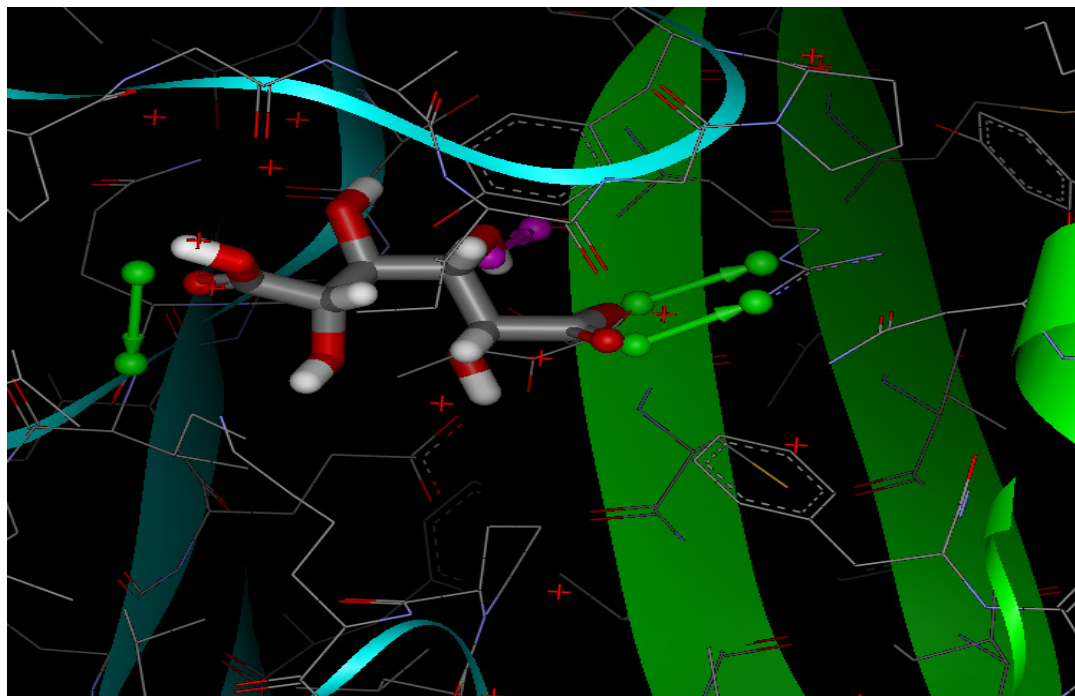


Figure 6.6 showing the active site of the protein in web lab preview, with Gluconic acid as the ligand attached, with unoptimized position of the left hand hydrogen bond acceptor

### 6.2.4 Pharma *TdcF* 4

Based on the orientation of hits obtained, slight changes were carried out in this pharmacophore. The position of H-bond acceptor vector was moved forward so that the blob of the vector arrow head touched the peptide nitrogen of the ILE33 and GLN32 (Figure 6.8). Additional constraint was added in the form of H-bond donor vector towards the carboxyl oxygen of the GLY31. Positions of the rest of the vectors were kept the same (Figure 6.7). Database search gave 53 hits in exactly 53 minutes.

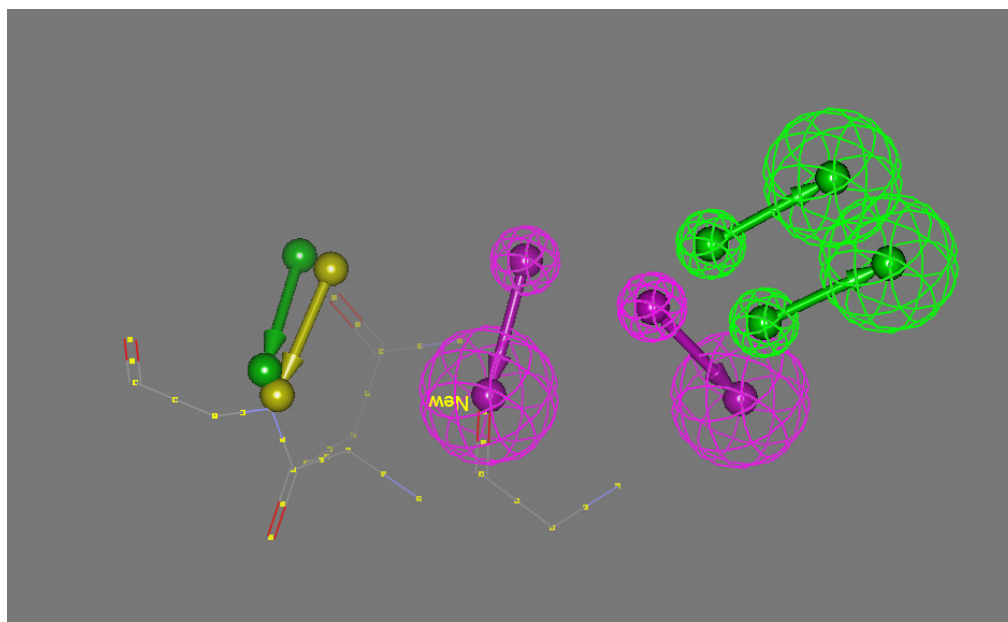


Figure 6.7 Pharmacophore with additional additional H-bond donor (labelled new) towards the carboxyl oxygen of GLY31, the position of H-bond acceptor (highlighted in yellow) moved more towards the peptide nitrogen of ILE33 and GLN32 (uncertainty spheres are removed for clarity viewing).

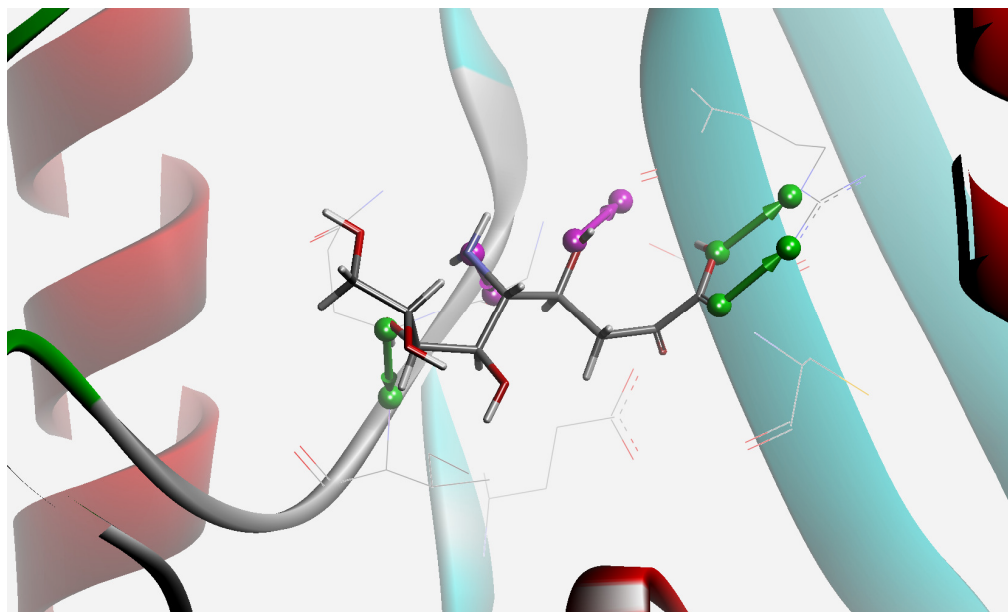


Figure 6.8 view of Protein binding site in web lab preview, showing Neuraminic acid with optimized position of the hydrogen bond acceptor vector (green) towards the left hand side

### 6.2.5 Pharma *TdcF* 5

The hits obtained from previous pharmacophore were not satisfying the pharmacophore constraints, therefore the two H-bond donor and one H-bond acceptor (towards ILE33 and GLN32) were removed. By looking at the binding mode in the co-crystallized structure {145} and the 3D structure of the protein a new constraint in the form of H-bond acceptor vector was introduced towards the CYS107 residue.

### 6.2.6 Pharma *TdcF* 6

On the basis of orientation of maximum hits obtained from previous pharmacophore, an H-bond donor vector was placed pointing towards the carboxyl oxygen of GLY31. H-bond acceptor vector towards the CYS107 was removed. Another H-bond acceptor vector pointing from the nearby crystallized water molecule towards the peptide nitrogen of ILE33 and GLN32 was introduced.

### 6.2.7 Pharma *TdcF* 7

Hits from previous pharmacophore showed that the H-bond acceptor vector pointing towards the ILE33 and GLN32 is not of critical importance therefore was

removed. H-bond acceptor vector was introduced towards the peptide nitrogen of the CYS107. The two H-bond acceptor vectors towards the guanidinium group of ARG105 were kept at the same position.

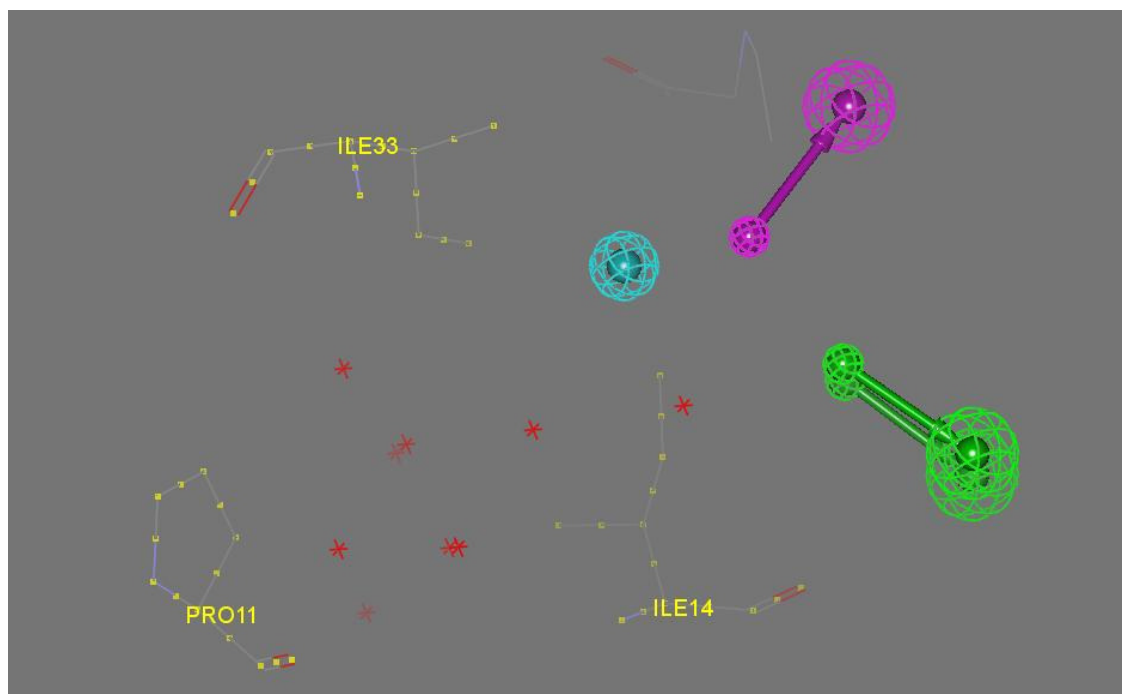
### **6.2.8 Pharma *TdcF* 8**

This time instead of using Cerius 2 and weblab viewer pro the pharmacophore was generated by using the DSV, Vector method as described in section 3.5.2. A total of 3 constraints were introduced in this pharmacophore. To increase the stringency on the constraints the radius of uncertainty/location spheres was decreased from 1.2Å to 0.8Å for the H-bond vector arrow heads and from 0.4Å to 0.35Å for H-bond vector arrow tail. H-bond acceptor pointing towards the CYS107 was removed and an H-bond donor vector was introduced pointing towards the carboxyl oxygen of the CYS107.

### **6.2.9 Pharma *TdcF* 9**

This pharmacophore comprised of 4 constraints, 3 constraints were the same as in the previous pharmacophore. An additional constraint in the form of hydrophobic point was introduced in the hydrophobic pocket of the binding site which is mostly occupied by the crystallized water molecules (Figure 6.9).





**Figure 6.9 Pharmacophore with additional Hydrophobic point (blue blob) in the hydrophobic pocket mostly occupied by water molecules (red stars), the amino acid residues ILE33, ILE14 and PRO11 can be seen (highlighted with yellow dots) (exclusion spheres are removed for clarity).**

Initial screening of *TdcF* as a search model gave a number of convincing hits which fit the active site cleft. In comparison to other ligands Neuraminic acid seemed to make interactions with the best geometry and filled the active site cleft. This compound with other known ligands was tested experimentally for binding studies.

**Table 6.1 Various pharmacophore optimization stages, the corresponding number of hits, and the database search time in minutes for *TdcF* (For optimization stage 1-8 Naturalism database was used, while for optimization stage 9-13 Che EBI database was used, similarly from optimization stage 1-7 Cerius 2 and Weblab viewer pro softwares were used, while for optimization stage 8-13 DSV® was used.**

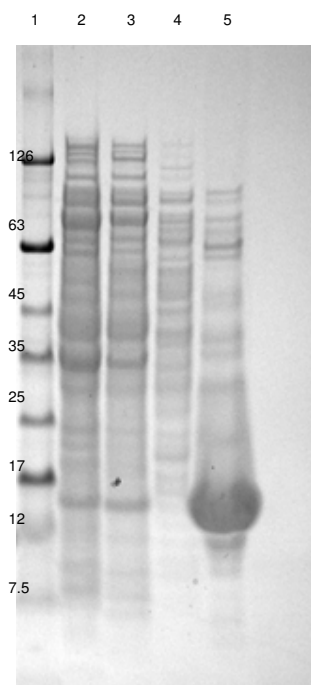
Optimization stage	Operation parameters	Number of hits	search time (minutes)
1	3 hydrogen bond acceptor vectors, 1 hydrogen bond donor vector	178	123
2	Change in size of uncertainty spheres, Change in position of H-bond donor	65	27
3	Change in position and length of H-bond vectors	88	77
4	Addition of H-bond donor vector and slight change in position of one of H-bond acceptor	53	53
5	Removal of two H-bond donors and one H-bond acceptor along with the addition of a H-bond acceptor at a new point	300	12
6	H-bond donor towards GLY31 and H-bond acceptor towards GLN32 and ILE33	1	1
7	Removal of H-bond acceptor from GLN32 and introducing towards CYS107	300	2
8	Removal of H-bond acceptor and addition of H-bond donor towards CYS107, along with decrease in radius of the uncertainty spheres around the head and tail of the vectors	300	26
9	Addition of constraint in the form of hydrophobic point	50	77
10	Introduction of query feature ligand with multiple atoms assigned to the same query atom	234	4
11	Specifying C2-O3 contact as double bond in the query	95	3
12	specifying C2-C3 contact as double bond and C2-O3 contact as single bond along with the addition of H-atom to O1 and O3,	6	3
13	Changing C2-C3 contact as single bond and removal of H-atoms from O1 and O3	240	3



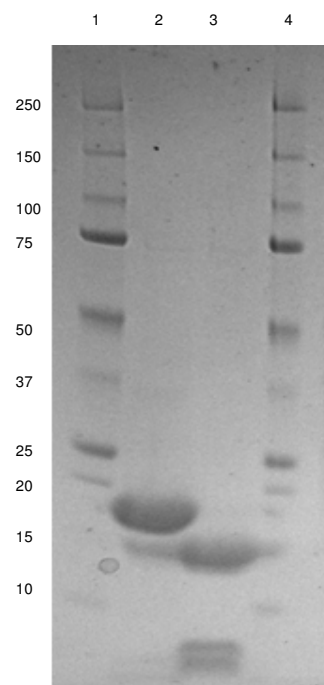
### 6.3 Expression and Purification of *TdcF*:

At the same time as we performed the in-silico screening of *TdcF*, we requested the plasmid for the protein from Prof.Dr.Gary Sawers, Martin-Luther University Halle, Germany. The requested plasmid of the protein was received later on. This allowed the purification of the protein to permit the assessment of a number of biophysical techniques for measuring the ligand binding.

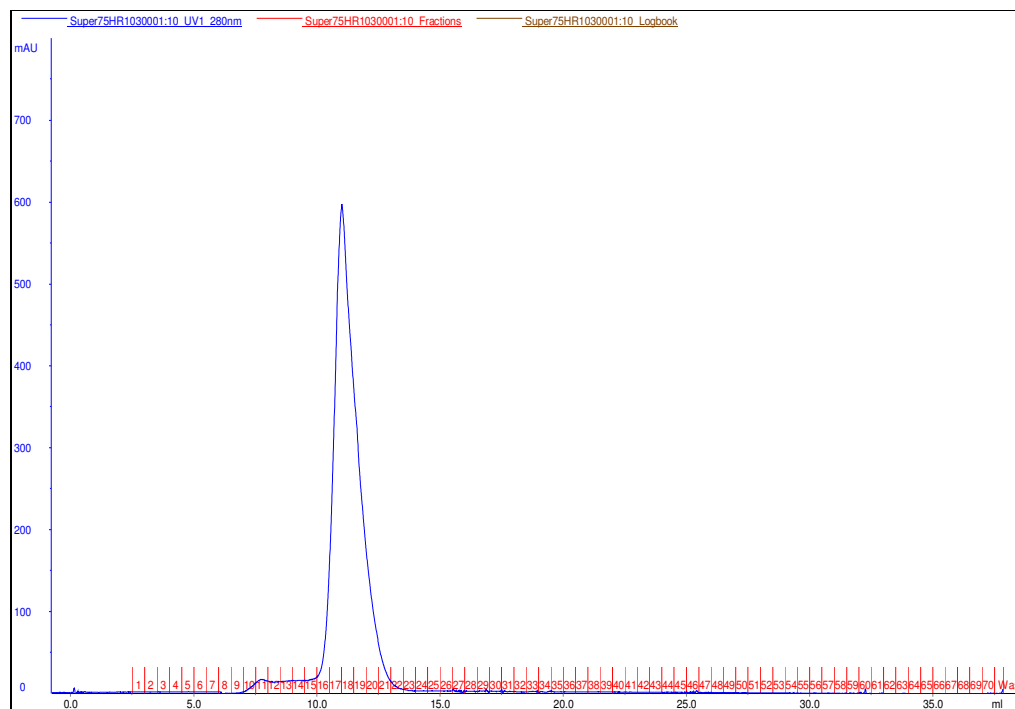
The pET-FG plasmid for the protein was transformed in to BL21 (DE3) cells and expressed in auto induction media (for co-crystallization purposes) and in <sup>15</sup>N labelled M9 media (for NMR experiments). The N-terminal His tagged protein was purified by using Nickel affinity chromatography (Figure 6.10). By using Amicon® concentrator (with a MWCO of 5KDa) the protein was simultaneously concentrated and buffer exchange in to cleavage buffer (20mM Tris-HCl, 150mM NaCl pH: 7.5) to get rid of imidazole. To remove the Histidine tag thrombin cleavage was carried out by adding thrombin to the protein as per manual (Thrombin kits, User protocol TB188, Novagen). The protein was then purified by passing through benzamidine sepharose column to get rid of thrombin. Thrombin cleavage was confirmed by running SDS-PAGE on pre and post cleaved samples (Figure 6.11). The final purification step was attained by loading the protein onto sephadex 75 gel filtration column using AKTA® (Figure 6.12)



**Figure 6.10** SDS gel image for TdcF after passing through Nickel column. Key: 1 = Marker, 2 = Flow through, 3 = Wash1, 4 = Wash2, 5 = Elute



**Figure 6.11** SDS gel image for TdcF after thrombin cleavage Key: 1 = Marker, 2 = Protein with His Tag, 3 = Protein without His Tag 4 = Marker



**Figure 6.12** Chromatogram image for TdcF after passing through Sephadex 75 gel filtration column.

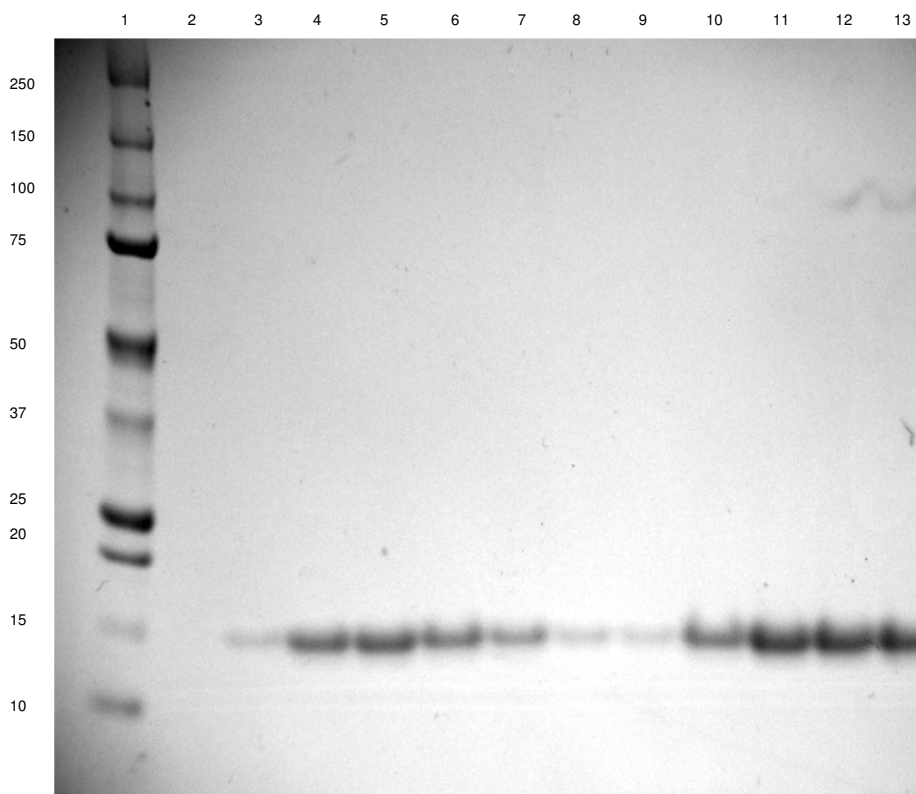


Figure 6.13 SDS gel image for *TdcF* fractions obtained from gel filtration chromatography (1 = marker, 2-13 = *TdcF* eluted fractions)

## 6.4 Ligand binding experiments against known ligands

### 6.4.1 DSC Experiments

Differential Scanning Calorimetry (DSC) experiments were carried out for *TdcF* against different ligands to detect binding. Normally binding can be measured using Isothermal titration calorimetry (ITC) but for weak binding ( $K_d$ , dissociation constant values in the micromolar to millimolar range) the ITC method is not particularly sensitive and requires significant amounts of protein to obtain any results. DSC experiments were carried out with the expectation that ligand binding will stabilize the thermal unfolding of the protein and thus resulting in an increase in  $T_m$  (transition mid point temperature).

The protein and ligand samples were degassed prior to loading. The rate of increase of temperature was 1°C/minute, with a temperature range from 15°C-100°C. The concentration of protein used was 71.4µM (1mg/mL) and the concentration of ligands used was 1mM. The ligands used were Pyruvate and Lactate. As *TdcF* structure contains two cysteine residues per monomer, one

partially buried and the 2<sup>nd</sup> one surface exposed therefore Dithiothreitol (DTT) was added to the buffer to prevent the formation of disulphide bonds between the cysteine residues. Protein aggregation problems were encountered when performing the experiments with the presence of a reducing agent in the buffer. In the beginning the presence of DTT in the buffer (20mM Tris, 0.5mM DTT, pH: 7.8) caused aggregation problems. The resultant DSC scans for the protein, both with and without ligand were irregular (Figure 6.14)

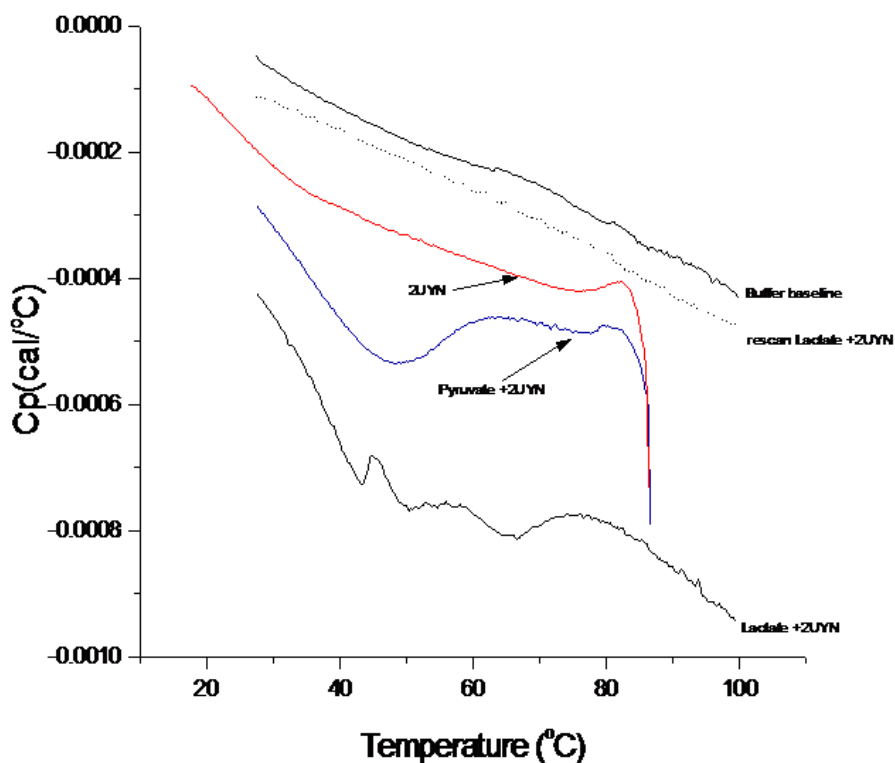


Figure 6.14 Abnormal DSC graph in the presence of DTT in the buffer, the irregular pattern of 2UYN (*TdcF*) can be observed both with and without ligands

Later on with the removal of DTT from the buffer normal DSC scans were obtained. The DSC scans of protein with and without DTT showed that the  $T_m$  of the protein in the presence of DTT is around  $85^\circ\text{C}$ , while in the absence of DTT the  $T_m$  goes down to around  $65^\circ\text{C}$  which might be because of the binding of DTT to the protein and stabilizing the native fold. In case of ligands, the addition of both pyruvate and lactate, no significant change in the  $T_m$  of the protein was observed (Figure 6.15). The DSC results suggested that either there is no binding at all or very weak to be observed through DSC experiments.

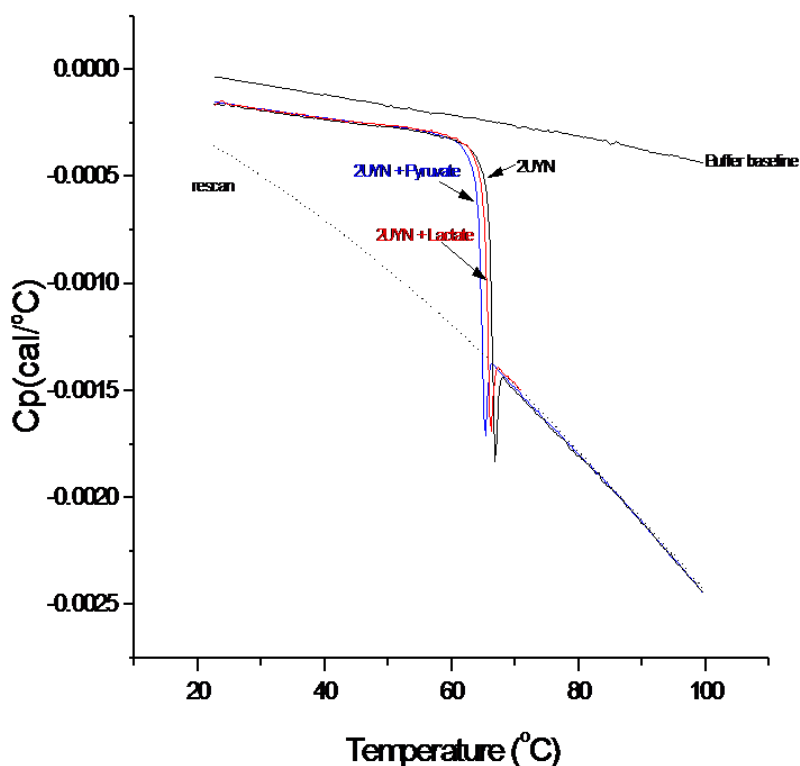


Figure 6.15 Normal DSC scans for 2UYN (*TdcF*) in the absence of DTT in the buffer, no significant changes in  $T_m$  are visible after the addition of different ligands (Pyruvate, Lactate).

## 6.5 Circular Dichroism (CD) Experiments

The results obtained from DSC were not helpful in identifying the binders therefore CD experiments were carried out in the near and far UV region. The CD spectrum for the native protein gave clear sharp peaks in the near UV region (250nm-300nm). The near UV spectrum (Figure 6.16) showed the secondary structure of the protein corresponding to the aromatic amino acids like tryptophans (285nm-295nm), tyrosines (275nm-285nm) and phenylalanine (265nm-275nm). The CD experiments were carried out to test if the addition of various ligands causes any significant changes in the secondary structure of the protein. The ligands used were N-acetyl neuraminic acid, threonine, lactate, alanine, pyruvate and serine. No significant shifts were observed in the near UV spectra of protein upon the addition of the ligand (Figure 6.17) except pyruvate which conferred a slight shift in the spectra (Figure 6.18). The subtle change in the near UV spectrum (250-280nm) brought about by the addition of pyruvate might suggest a plausible change in the aromatic environment of the protein. The concentration of protein used in near UV and far UV CD experiments was 96 $\mu$ M and 24 $\mu$ M respectively. The path length of the cell used for near UV and far UV experiments was 0.5cm and 0.02cm respectively. 20mM Tris, 0.5mM DTT pH: 8.0 buffer was used through out the experiments. The ligands stock solutions were prepared in the same buffer and the pH was adjusted to 8.0

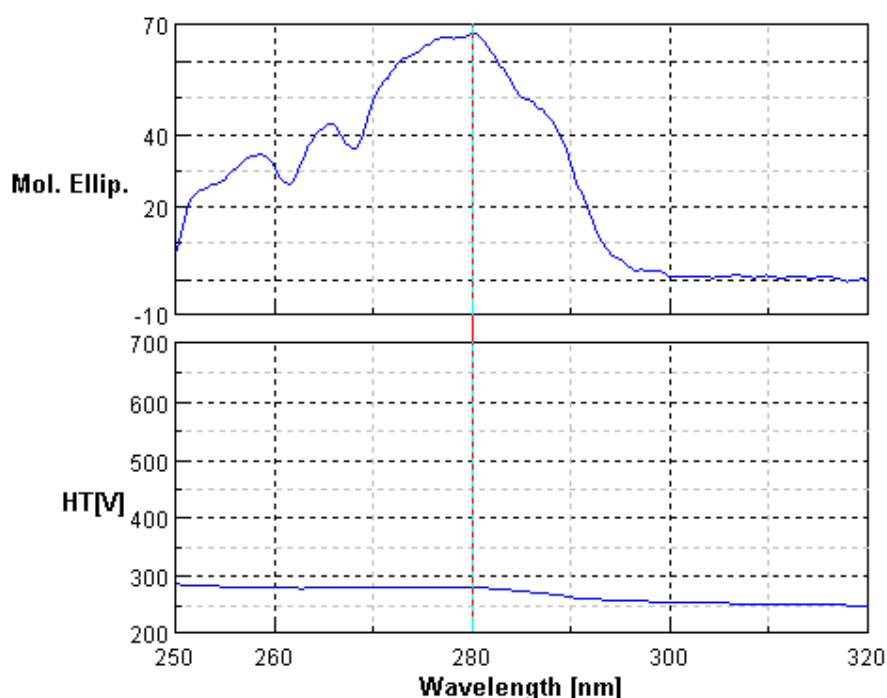


Figure 6.16 Near UV CD spectrum for *TdcF* alone

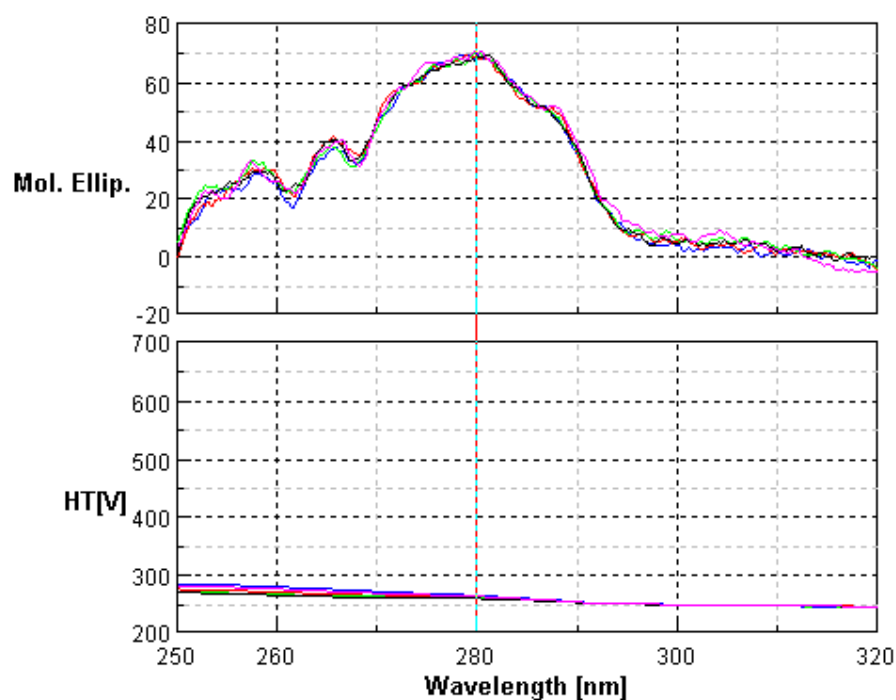


Figure 6.17 Overlaid near UV CD spectra of *TdcF* in the absence (blue) and presence of N-acetyl neuraminic acid (green), DL-Threonine (black), L-lactate (pink) and DL-alanine (red)

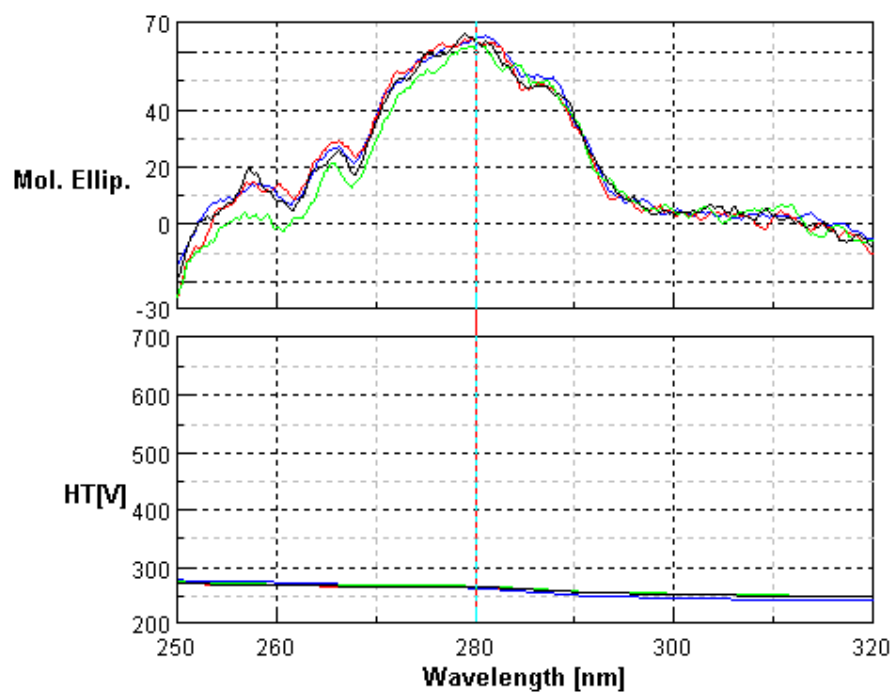


Figure 6.18 Overlaid near UV CD spectra for *TdcF* in the absence (blue) and presence of N-acetyl neuraminic acid (red), Pyruvate (green) and DL-Serine (black), slight shift with pyruvate can be observed in the 250-280nm range of the spectrum

The Far UV spectra for the protein gave clear double minima around 208nm-225nm corresponding to the  $\alpha$ -helices in the protein. The addition of ligands caused no significant change in the far UV spectra (Figure 6.19 & 6.20)

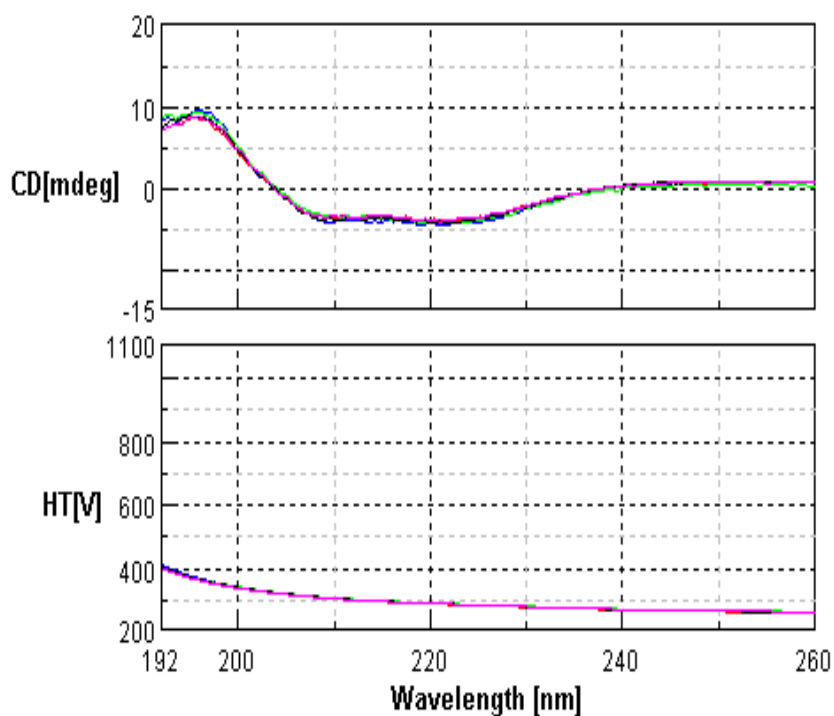


Figure 6.19 Overlaid far UV CD spectra for *TdcF* in the absence (blue) and presence of N-acetyl neuraminic acid (green), DL-threonine (black), L-lactate (pink) and DL-alanine (red)

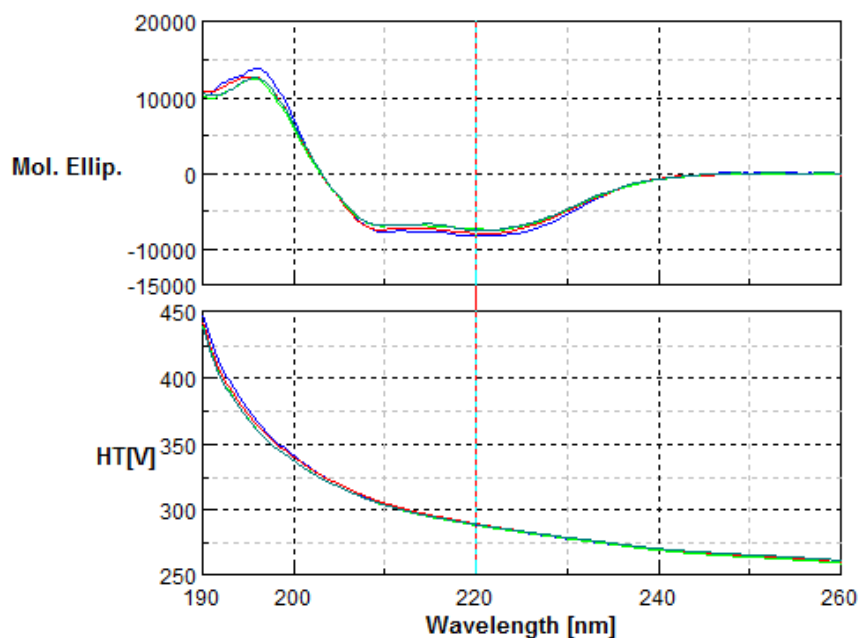


Figure 6.20 Overlaid far UV CD spectra for *TdcF* in the absence (blue) and presence of L-lactate (green), L-serine (pink) and pyruvate (red), Protein = 24 $\mu$ M, ligand = 454 $\mu$ M).



## 6.6 NMR experiments

The CD experiment results were not very encouraging to put light on the ligand addition effect therefore protein was further used for NMR (Nuclear Magnetic Resonance) studies. The protein ligand binding data was gathered by carrying out NMR titrations via HSQC (Hetero Nuclear Single Quantum Correlation) experiments.  $K_d$  values were determined for different ligands based on difference in the chemical shift values.

### 6.6.1 NMR titrations

Samples of pure  $^{15}\text{N}$ -labelled protein with a concentration of 25mg/ml in 20mM sodium phosphate buffer pH 7.5 were used for NMR titrations. A Bruker Advance 600 MHz spectrometer equipped with triple resonance cryoprobe and gradient channel was used to record a series of  $^{15}\text{N}$ -HSQC spectra. NMR sample for loading onto the spectrometer was prepared in the NMR tube by the following ratio:

$$570\mu\text{L (protein)} + 30\mu\text{LD}_2\text{O} + 1\mu\text{L DTT (1M)} = 601\mu\text{L}$$

The extent of binding of ligands to *TdcF* was determined by monitoring the change in chemical shifts of HSQC cross-peaks as a function of the ligand concentration. To do this, a 0.71 mM solution of  $^{15}\text{N}$ -labelled *TdcF* in 20 mM sodium phosphate, 150 mM NaCl, and 1 mM DTT pH 7.5 buffer was titrated against the same sample containing ligand. Ligand addition was carried out by mixing pure protein sample with sample containing protein and maximum ligand. The formula used to calculate the exchange volume is given below.

$$\text{Volume to exchange} = \text{total volume} \times \frac{(\text{desired ratio} - \text{low ratio})}{(\text{high ratio} - \text{low ratio})}$$

The concentration of protein was confirmed through UV/Vis spectroscopy. The ligand solution was prepared in the same buffer as for the protein and the pH was adjusted.

### 6.6.2 HSQC Spectra and $K_d$ determination

In a protein structure different Protons (H) attached to Nitrogen (N) in different amino acid residues have different chemical environment in the protein. They may be in close proximity to a hydrophobic residue or to a hydrophilic residue. Due to this specific chemical environment around each amide nitrogen and amide proton (corresponding to a specific amino acid residue) they acquire a specific chemical shift value which in turns leads to the appearance of individual peaks in a HSQC spectrum. The chemical shift values for specific amino acids can change due to the change in chemical environment around these residues. The change in chemical environment can be due to pH changes (causing protonation or deprotonation of certain amino acid residues) or due to the addition of ligand (causing change in the chemical environment or by direct contact). The spectra were processed by picking the cross peaks for amide nitrogen and amide proton of individual amino acid residues and then assigning chemical shift values to them. The changes in position of individual cross peaks were followed upon addition of ligand by using the *CcpNmr Analysis* v.2. program {146}. The changes were measured in the form of chemical shift differences. The values obtained from chemical shift differences were then entered in to a nonlinear fit (Eq: 1). The equation related the resonances in fast exchange between free and bound protein. The quadratic equation in terms of non-linear fit {147} used to calculate the  $K_d$  values for individual ligands is given below

$$\Delta\Omega_i = \frac{K_d + [L_T] + [E_T] - \sqrt{\{(K_d + [L_T] + [E_T])^2 - 4 [L_T] [E_T]\}}}{2[E_T]} \Delta\Omega \quad \{ \text{Eq: 1} \}$$

Where  $[E_T]$  and  $[L_T]$  are total concentrations of enzyme and ligand added,  $\Delta\Omega_i$  and  $\Delta\Omega$  are the chemical-shift differences between the observed and initial and final and initial resonances, respectively. The more simpler mathematical form of Eq:1 can be given in the form of Eq:2 as:

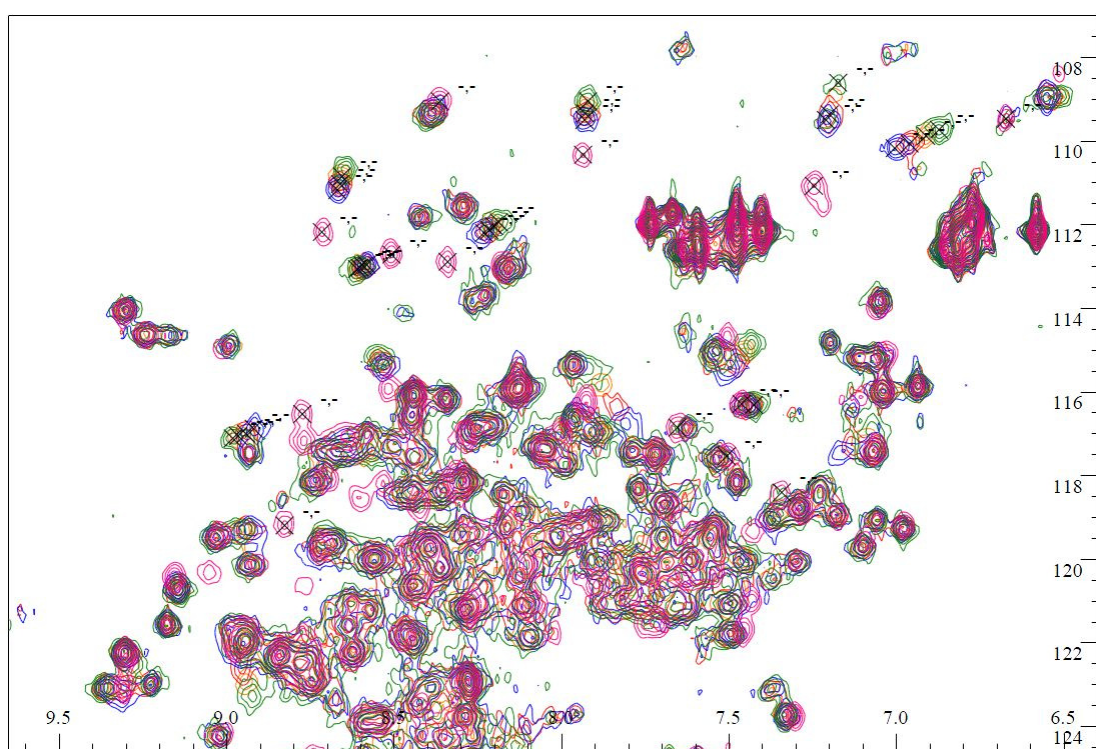
$$Y = A \{ (B+X) - \sqrt{\{(B+X)^2 - 4X\}} \} \quad \{ \text{Eq: 2} \}$$

Where Y= Chemicals shift change, A = maximum Chemicals shift change/2

$$B = 1 + K_d / [\text{protein}], X = \text{Ligand/Protein}$$

### 6.6.2.1 *TdcF* with 2-Ketobutyrate

In the crystal structure of the protein it can be observed that the enol form (double bond between C2 and C3 and hydroxyl group at C2) of 2-Ketobutyrate binds in the binding site of the protein {145}. This indicates that 2-ketobutyrate fits in a more well defined manner within the binding site by satisfying all the H-bond interactions including the hydrophilic and hydrophobic interactions. The  $K_d$  value calculated for 2-ketobutyrate was 204 $\mu$ M which is the lowest among all the other ligands. The tighter binding of 2-ketobutyrate indicates that the protein prefers the enol form on the keto form of the potential ligands.



**Figure 6.21** A series of overlaid  $^{15}\text{N}$ -HSQC spectra for *TdcF* with increasing 2-ketobutyrate concentration (Pink spectra is for protein alone, blue, red, orange and green spectra are at 4, 8, 17 and 25mM ligand Concentration respectively).

By observing the crystallization conditions for *TdcF* it appeared that the protein was crystallized at pH 8.0 {145}. It has been experimentally demonstrated that the pH is a major factor in altering the keto/enol equilibrium of pyruvate. The ratio of keto/“enol” becomes unity at pH 5.8 and decreases with increase in pH and vice versa; e. g., the keto form is approximately 10 at PH 4.8 and 0.1 at pH 6.8 there fore it can be estimated that at pH 8 the keto form will be 0.01 {148}.

The pH crystallization conditions {145} and the pH dependant keto/enol equilibrium experiments {148} indicate that there is a higher proportion of enol form than the keto form at pH 8.0. Thus it gives a plausible explanation of tighter binding of 2-ketobutyrate (enol form) (Figure 6.21).

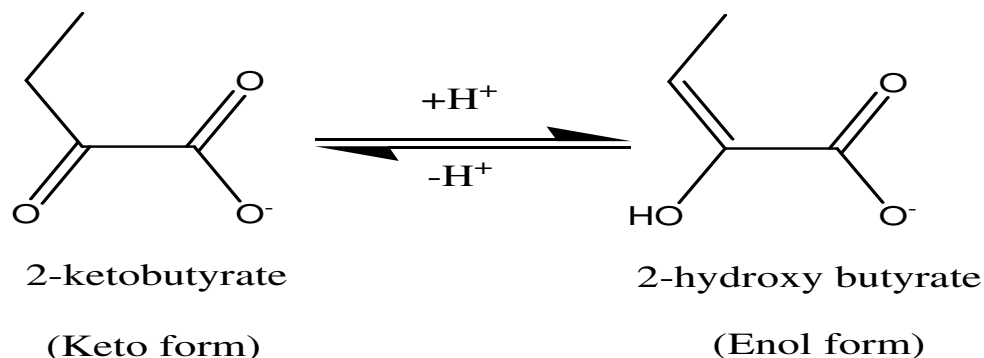
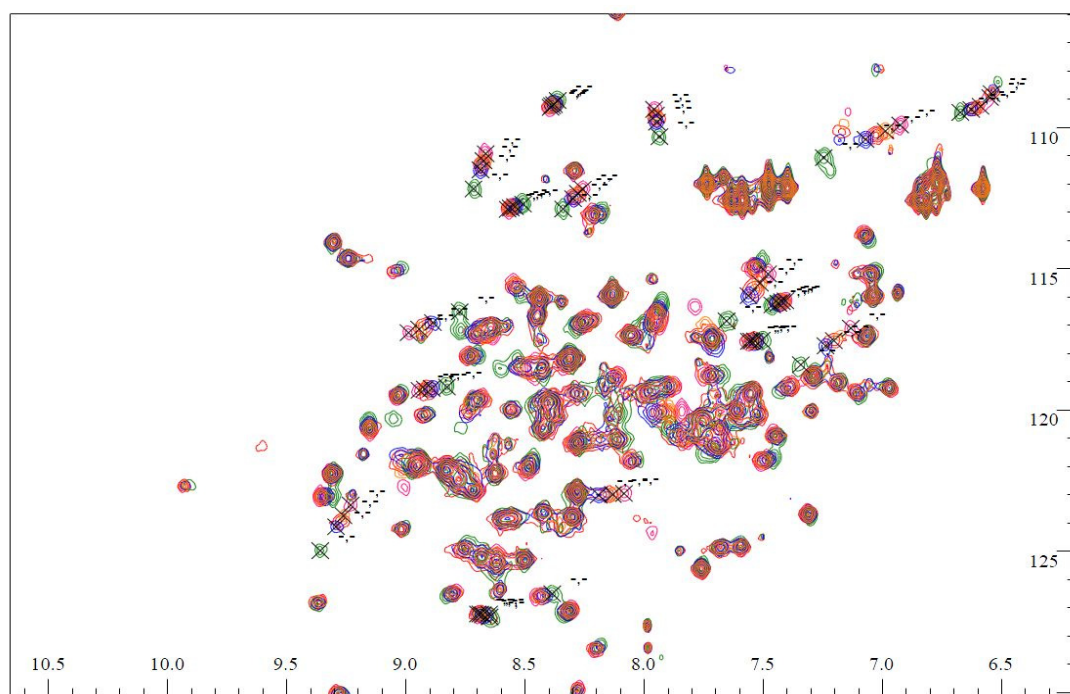


Figure 6.22 Equilibrium tautomers of the 2-ketobutyrate in keto and enol form

#### 6.6.2.2 *Tdcf* with pyruvate

The  $K_d$  value calculated for pyruvate was  $7000\mu\text{M}$ . This indicated a 35 times weaker binding than 2-ketobutyrate ( $K_d = 204\mu\text{M}$ ). It has been experimentally determined that both keto and enol forms of pyruvate can exist at equilibrium and at a pH of 5.8 the ratio is unity and with increase in pH the enol form dominates {148}. The weak binding of pyruvate in comparison to 2-ketobutyrate can be associated with the protein preference to bind to the keto form of pyruvate instead of the enol form and thus shifting the equilibrium towards the keto form at pH 8.0.



**Figure 6.23** A series of overlaid  $^{15}\text{N}$ -HSQC spectra for *TdcF* with increasing Pyruvate concentration (Green spectra is for protein alone, blue, red, orange and Pink spectra are at 3, 5, 7.5 and 12.5mM ligand Concentration respectively).

### 6.6.2.3 *Tdcf* with L-Lactate

The  $K_d$  value calculated for L-Lactate was  $>25.0\text{mM}$ . The reason for not getting an exact value for L-Lactate was unsaturation of the binding curve. With the corresponding addition of 2-Ketobutyrate the shift distances for certain residues change and a point is reached at maximum ligand concentration when no change in shift distance is observed this is called the saturation point. In case of L-Lactate at the highest concentration the binding curve did not show the saturation stage (Figure 6.24). Thus the calculated  $K_d$  value was  $> 25.0\text{mM}$ .

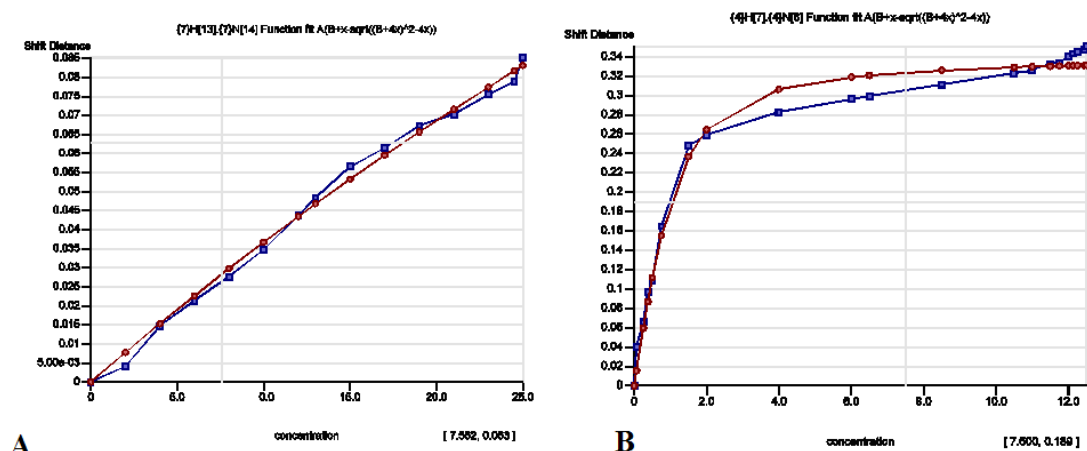


Figure 6.24 Binding curve for a single amino acid residue representing change in the chemical shift distance of a single amide cross peak due to the addition of ligand (A = L-Lactate, B = 2-ketobutyrate), the saturation stage is absent in A but visible in B, red curve = function fit, blue curve = data fit.

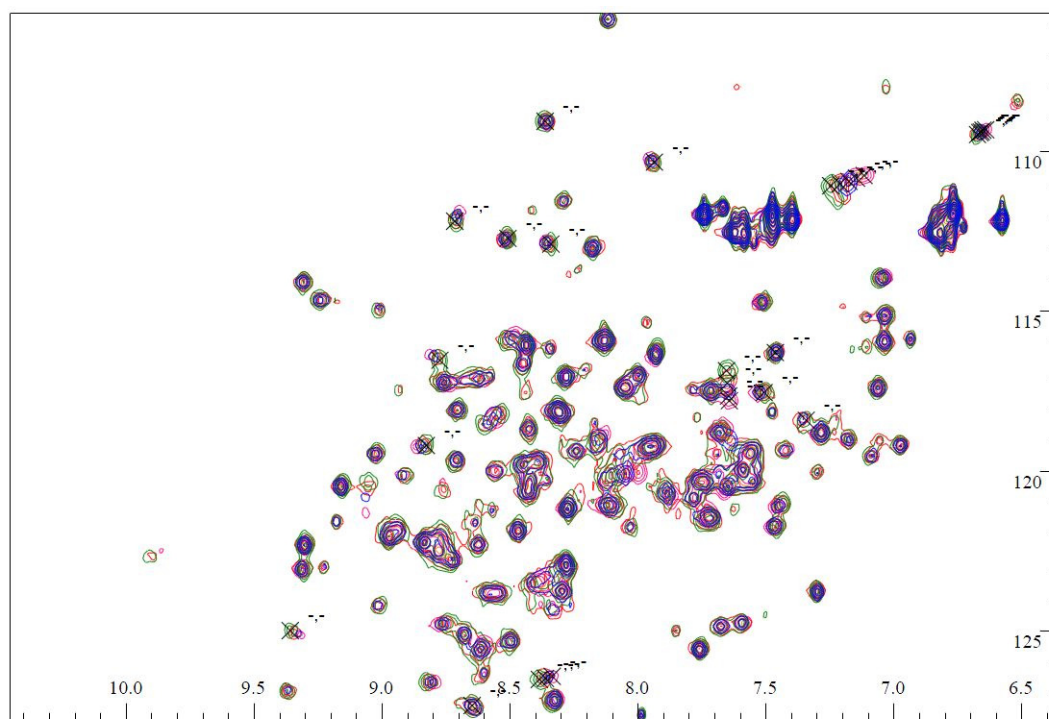
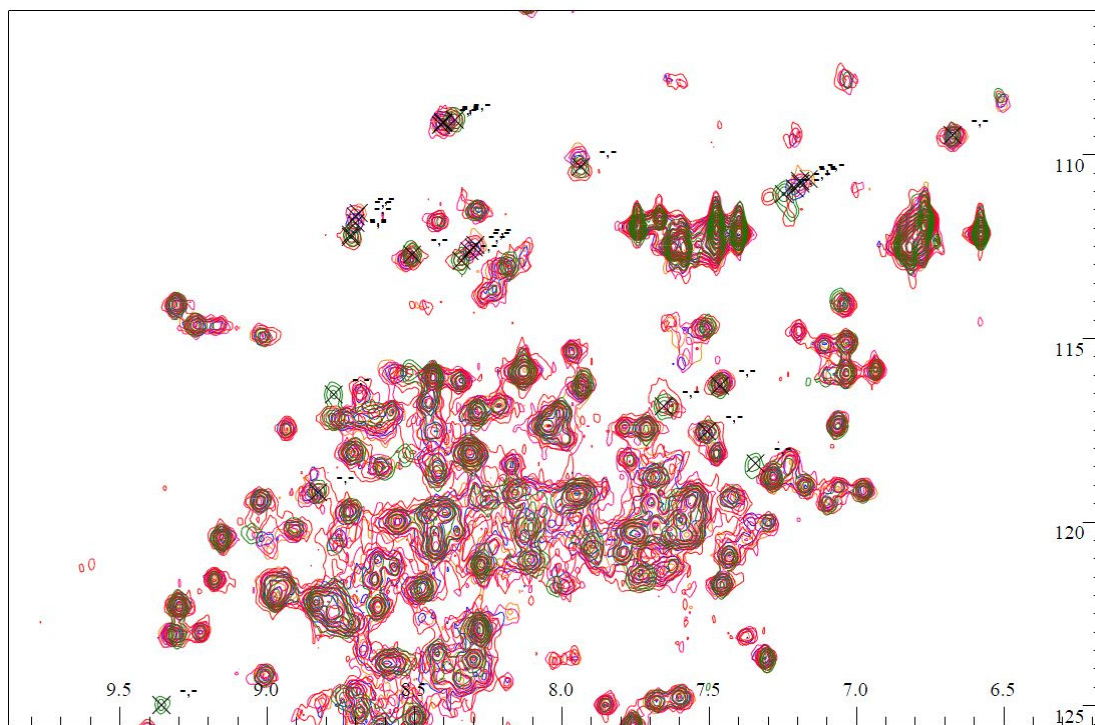


Figure 6.25 A series of overlaid <sup>15</sup>N-HSQC spectra for TdcF with increasing L-lactate concentration (Green spectra is for protein alone, blue, orange, pink and red spectra are at 10, 15, 20 and 25mM ligand Concentration respectively).

#### 6.6.2.4 Tdcf with L-serine

The crystal structure of TdcF with L-Serine indicates additional H-bond interactions between the amino group of L-Serine and carbonyl oxygen of ARG105 {145}. The  $K_d$  value calculated for L-Serine was 8.4mM. The  $K_d$  value for

L-Serine indicated that it has about the same binding affinity for the protein as Pyruvate ( $K_d = 7.0\text{mM}$ ) but binds more tightly than L-Lactate ( $K_d = > 25.0\text{mM}$ ).

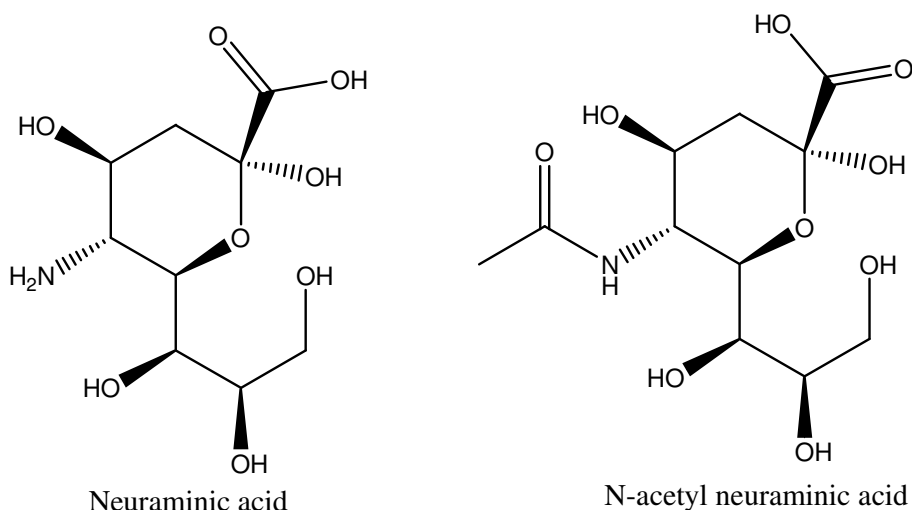


**Figure 6.26** A series of overlaid  $^{15}\text{N}$ -HSQC spectra for *TdcF* with increasing L-Serine concentration (Green spectra is for protein alone, red, blue, pink and orange spectra are at 4, 8, 10 and 12mM ligand Concentration respectively)

#### 6.6.2.5 *Tdcf* with N-acetyl Neuraminic acid

The pharmacophore hit Neuraminic acid (tried in the form of N-acetyl neuraminic acid) did not show any binding at all. There was no change in chemical shift, and the HSQC spectra for both the samples with and without the ligand appeared the same. There can be various possibilities of not binding, one of them could be the open and closed form of neuraminic acid. It depends upon the ratio of open form to closed form at equilibrium stage and then on the likely form to bind to the protein. If for instance open form is preferred for binding, but at equilibrium point the concentration of open form is much less than the closed form (1:200), then there are very less chances of binding of open form to the protein. Another reason which might suggest why no binding took place at all, could be because of using the acetylated form of neuraminic acid while the hit obtained from pharmacophore searching was neuraminic acid. Figure 6.27 shows the chemical difference between Neuraminic acid and N-acetyl neuraminic acid.





**Figure 6.27** Image showing chemical structure difference between Neuraminic acid and N-acetyl neuraminic acid.

### 6.6.3 $K_d$ values of different ligands for *TdcF*

$K_d$  values for different ligands in terms of weak and tight binding can be explained in the form of simple equation given below

$$K_d = [P] [L] / [C]$$

Where P = protein, L = ligand and C = protein-ligand complex. If at a given equilibrium stage [C] is bigger than [P] and [L], it gives a low  $K_d$  value. A low  $K_d$  value accounts for tight binding while a high  $K_d$  value corresponds to weak binding. A total of 5 ligands (N-acetyl neuraminic acid, 2 ketobutyrate, pyruvate, L-lactate and L-serine) were titrated against *TdcF* to calculate their  $K_d$  values. Table 6.2 gives a comparative account of ligand binding in terms of  $K_d$  values. NMR titrations revealed that some of the ligands bind very tightly to the protein and some very weakly. The binding affinity of different ligands for the protein in descending order is given as

**2-ketobutyrate > Pyruvate > L-Serine > L-Lactate**

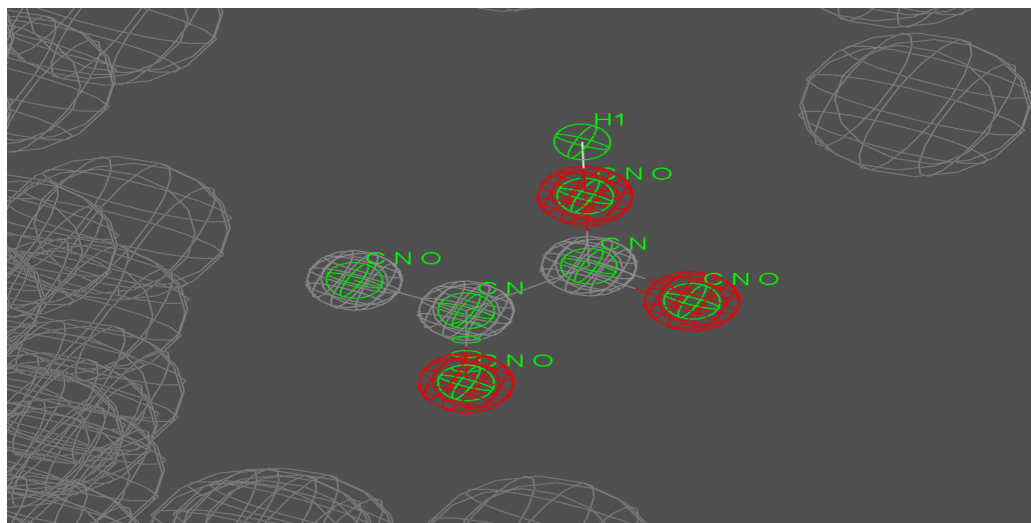
S.#	Ligand	Ligand concentration range (mM)	$K_d$ value (mM)
1.	2-Keto-butyrate	0-12.5	$0.204 \pm 0.01$
4.	Pyruvate	0-12.5	$7.3 \pm 0.03$
2.	L-Serine	0-25	$8.4 \pm 0.05$
3.	L-Lactate	0-25	>25.0

**Table 6.2** Different ligands and their concentration range along with calculated  $K_d$  values for *TdcF*



#### **6.6.4 Pharma TdcF 10 (Generation of Query ligand)**

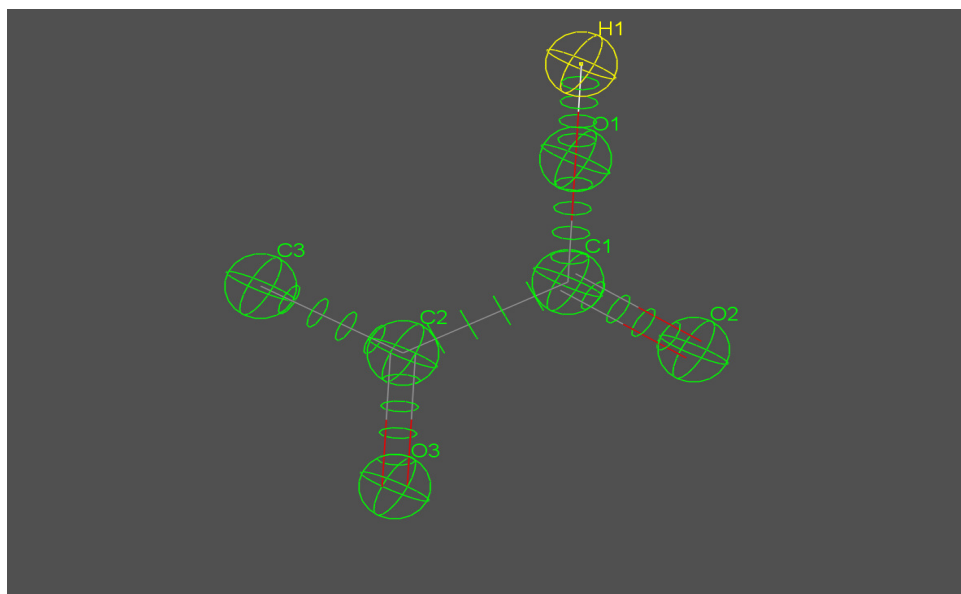
The  $K_d$  values obtained from NMR titrations and the co-crystallized structures demonstrated that the presence of carboxylate group on the ligand is important for recognition purpose within the binding site. With the privilege of having the X-ray structure of the protein with bound ligands and the Accelrys DSV® software with the feature of converting the ligand structure into query form. By using the Query atom method as described in section 3.5.3 a more sophisticated form of pharmacophore was generated. The Accelrys DSV® provided the option of flexibility in assigning multiple atoms to the same query atom along with attributing formal charge, partial charge and oxidation state to the query atoms. This assisted in the freedom of search through the data base. Different parameters for the query like bond type, charge on atoms and hydrogen count on each atom were specified by using Catalyst®. Initial search on the query was carried out by using the Catalyst® fast flexible results option. This showed whether the query gives hits or not. If in case the query did not give any hits then the query was further modified by optimizing different parameters until it gave hits. The final query was then subjected to database search. The location/size of uncertainty spheres around query atoms was set to 0.4Å. To avoid the steric clash between the ligand and the amino acid residues of the protein, exclusion spheres were introduced within a radius of 10.0Å around the center of the binding pocket (Figure 6.26). The query when searched against the naturalism database, provided 234 hits in just 4 minutes.



**Figure 6.28** Query feature pharmacophore generated through Accelrys DSV® with query ligand comprising of multiple atoms for the same query atoms (green spheres), surrounded by uncertainty/location spheres (small grey and red spheres), exclusion spheres are around the query pharmacophore (large grey spheres)

### 6.6.5 Pharma TdcF 11

To explore further the binding site and the type of hits, the query atoms were specified individually for each atomic position in the query and the bond between C2 and O3 was specified to be double (Figure 6.27). The number of hits obtained as a result decreased from 234 to 95. This showed that specifying the double bond at this position leads to significant loss of hits.



**Figure 6.29** Representation of query pharmacophore with individual atoms specified for each atomic position in the query ligand (the uncertainty spheres around atoms and exclusion spheres around the query ligand have been removed for clarity)

### 6.6.6 Pharma TdcF 12

As most of the hits obtained showed specific bonding between C2-C3 and C2-O3, therefore the query pharmacophore was modified by changing C2-C3 bond type from single to double and C2-O3 from double to single. The number of hits obtained as a result of this operation were only six and 2-hydroxy butyrate was absent among the hits.

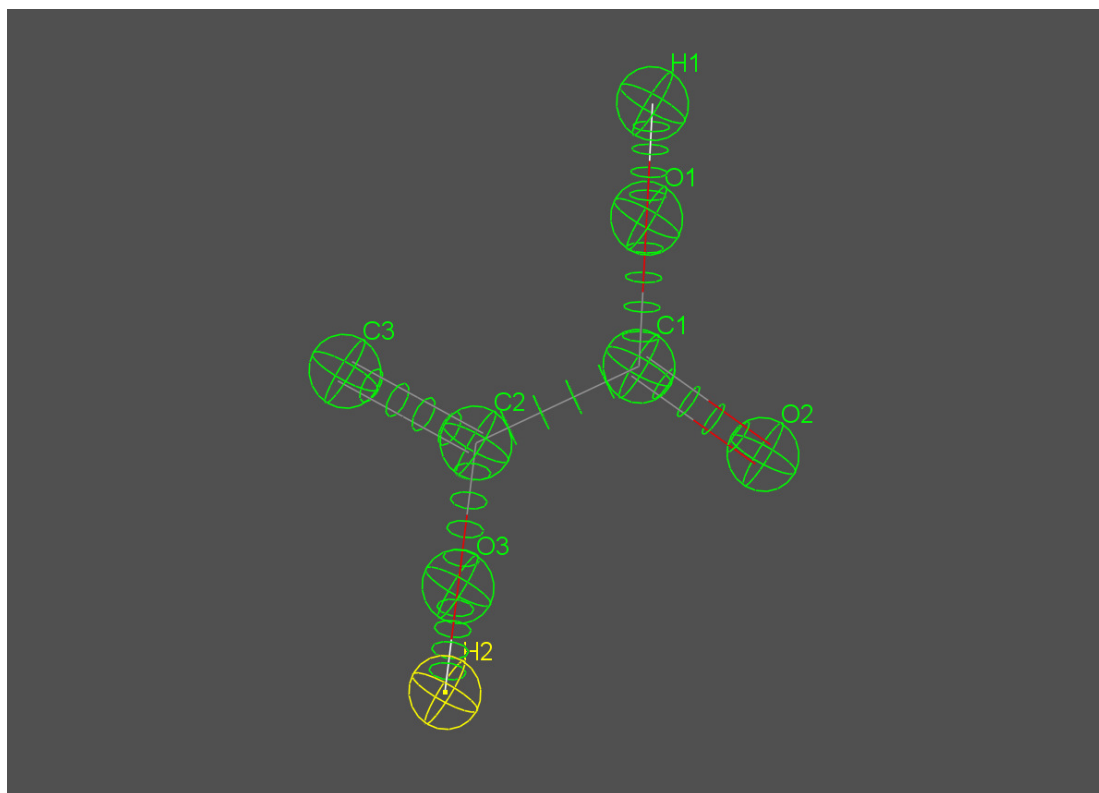
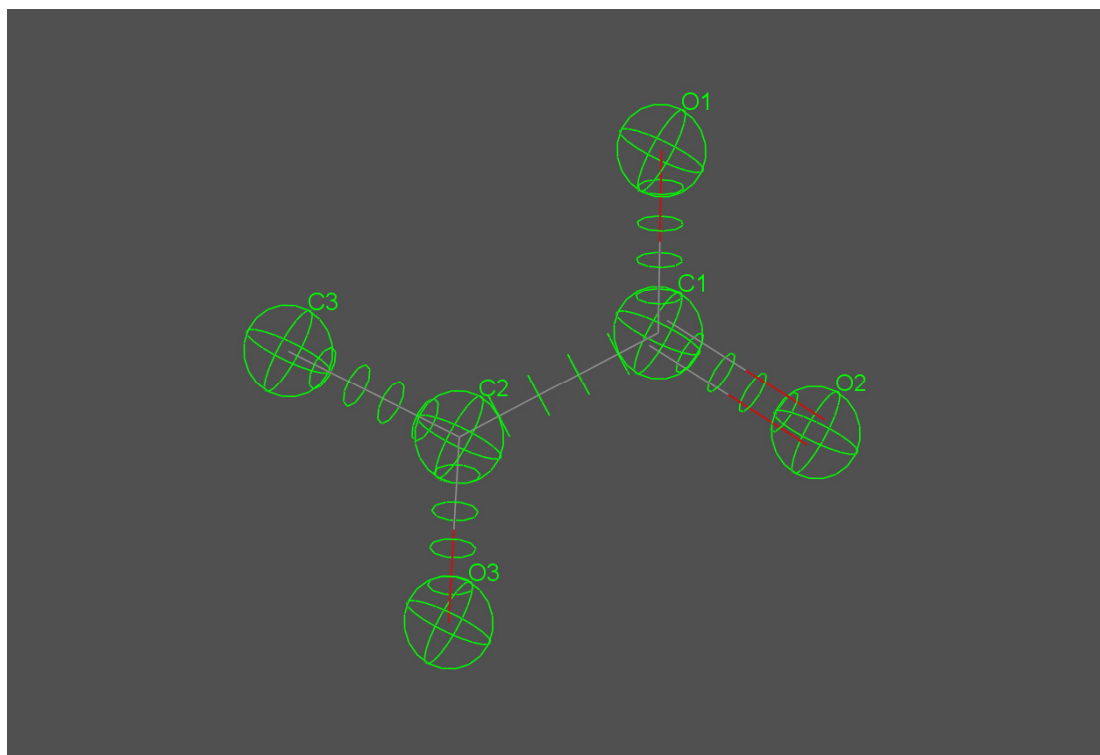


Figure 6.30 Representation of query pharmacophore with change in bond type from single to double between C2-C3 and from double to single between C2-O3.

### 6.6.7 Pharma TdcF 13

Addition of hydrogen atom on O1 and O3 and specifying the C2-C3 contact as a double bond in the last query pharmacophore caused losing of the 2-hydroxy butyric acid as a hit. Therefore in this pharmacophore model the hydrogen count on these atoms was set to free (any number) (Figure 6.29). This resulted in significant increase in number of hits and also regained the 2-hydroxy butyric

acid as a hit. The enol form of the 2-keto butyrate whose co-crystallized structure has been determined [145] was not in the database; instead the hydroxyl form was present which came up as a hit.



**Figure 6.31** Representation of pharmacophore in the form of query ligand in Accelrys DSV® with hydrogen atoms removed from O1 and O3. the exclusion and uncertainty location spheres have been removed for clarity.

## 6.7 Co-Crystallizations for *TdcF*

To determine the binding mode of a given ligand the glycerol stock of the protein was dialyzed overnight in 50mM Tris pH 7.5. The Protein was concentrated to 25mg/ml by using Amicon® concentrator. Crystallization trials were carried out using conditions around those known to yield crystals from the work of Lawson *et al* [149]. The protein was used at a concentration of 10mg/ml. PEG (Polyethylene glycol) of Mr 1500, 2000, 3350 and 8000 were used as a precipitant in the 4%, 6%,8%,10%,12%,14% and 16% concentration range. Crystalline precipitates were seen at PEG 1500 in the concentration range of 4%, 6%, 8% and 10%. The above crystallization conditions were further optimized to get individual crystals. Most of the optimized conditions resulted in crystalline precipitates.

## 6.8 Conclusions and Future work

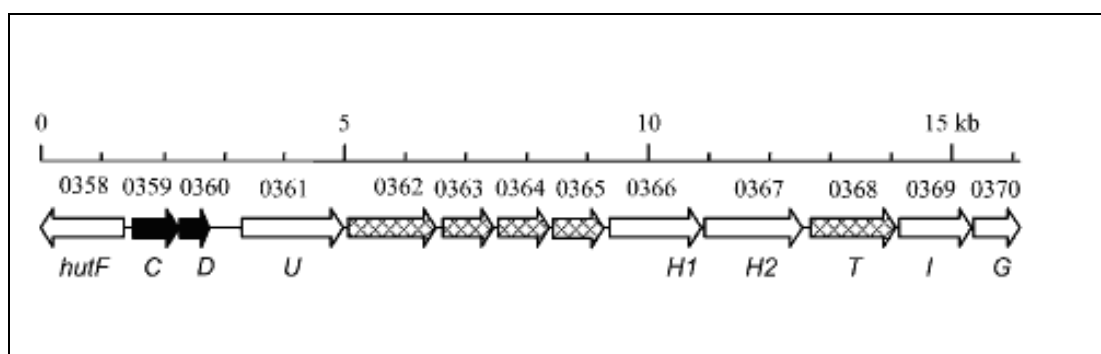
The  $K_d$  values obtained (Table 6.2) give good indications to the features required for tight binding. Namely compounds with carboxylate and a keto group at C2 position e.g; 2-ketobutyrate and pyruvate bind better than the ligands having tetrahedral geometry at C2 such as lactate. Serine has been shown to bind in a quite different conformation which accounts for its modest binding affinity.

- 1) Ligands having keto group at C2 bind preferentially to the protein.
- 2) 2-keto butyrate appears to be the best binder and could be a ligand at physiological concentrations (Table 6.2)
- 3) N-acetyl neuraminic acid does not bind to the protein but neuraminic acid should be tried to exclude this as a possibility.

To test if the ligand effect could be measured by CD, higher concentrations of ligands would need to be tried. NMR titrations showed the advantage of measuring weak and tight binding. TdcF was used as an initial test case for checking the compatability of pharmacophore searching for identification of suitable ligands. Hits obtained through pharmacophore searching included compounds like pyruvate and Lactate. Serine was also among the hit which has already been co-crystallized. The  $K_d$  values obtained for different ligands showed that 2-ketobutyrate is the best binder and could be the potential ligand of the protein at physiological concentration. The pharmacophore hit Neuraminic acid satisfies all the constraints in the pharmacophore and needs to be tested in its native form for binding purposes. The results from pharmacophore searching and the binding data in the form of  $K_d$  values suggests TdcF to be a binding protein which may be performing a regulatory role by binding to 2-ketobutyrate and preventing its accumulation. The results obtained by using Pharmacophore searching on TdcF are promising and further encourage its use on other unknown proteins to findout potential ligands for them. The potential ligands can be tested by using NMR for detecting binding in terms of  $K_d$  values.

## 7. HutD, a protein of unknown function from *Pseudomonas aeruginosa* PAO1

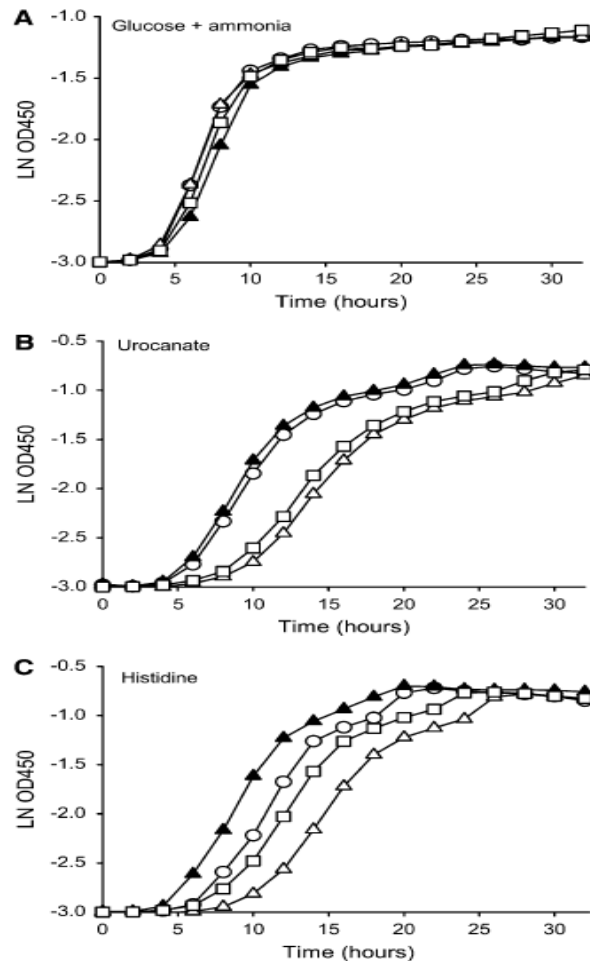
HutD (PA5104/PAD) is a conserved hypothetical protein of unknown function from *Pseudomonas aeruginosa* PAO1. Its gene is associated with the histidine utilization (Hut II) pathway operon. The hut genes are located in Hut operon with *hutD* located downstream of *hutC* in an overlapped manner (Figure 7.1). *HutD* encodes a protein of 200 amino acids with a Mr of 21.74 kDa that belongs to a group of uncharacterized proteins Pfam0596 that are highly conserved in bacteria. The crystal structure of HutD from *P.aeruginosa* PAO1 has been determined in our group and by others (PDB: 1YLL). The asymmetric unit contains four chains that are folded into four domains.



**Figure 7.1** Structure of the *P. fluorescens* SBW25 *hut* locus, the metabolic hut genes and the hut regulators are shown by open and solid arrows respectively, image adopted from {150}

HutD is a broadly conserved but functionally unknown component of the histidine utilization pathway in *P. auroginosa*. Recently, the *hutD* gene in a related organism *P. fluorescens* SBW25 has been genetically characterized and its involvement in histidine and urocanate utilization has been suggested by gene deletion studies {150}. In the study the growth of deletion mutants relative to wild type at different media composition was determined. It was observed that the *hutD* deletion mutant grows slowly on histidine or urocanate compared with the wild-type.  $\Delta hutC$  mutant and  $\Delta hutCD$  mutant showing that *hutD* functions independently of *hutC* (Figure 7.2).

On the basis of this data *hutD* has been suggested to be governing the activation of *hut* operon. The biological need for a regulator may relate to the potentially harmful effects of high rate of histidine catabolism, as one of the consequences resulting from the high levels of *hut* activity is the build-up of intracellular ammonium [150]. Although the structure of HutD is available it has proved little information into its function and the molecular mechanisms of HutD action remains unknown.



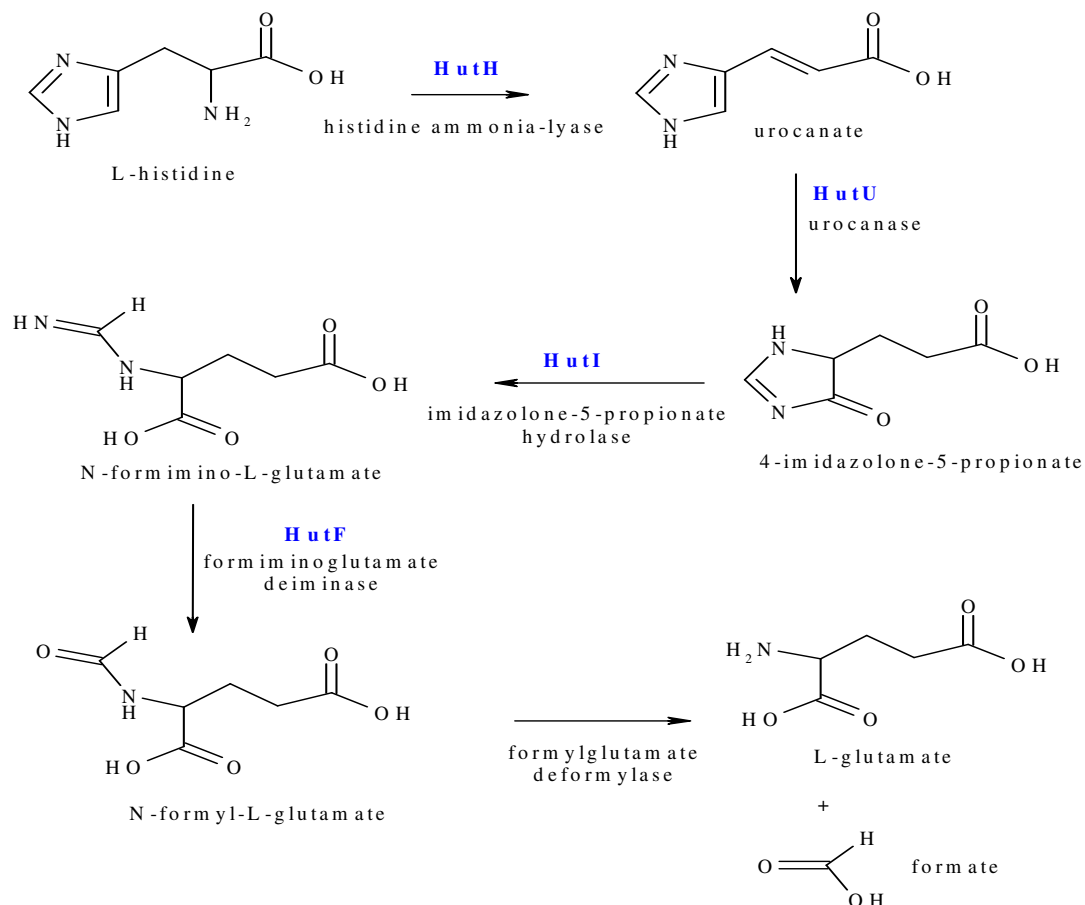
**Figure 7.2** Growth kinetics of the *hutCD* deletion mutants in different growth media compositions. wild-type SBW25 = open circles, mutant  $\Delta hutC$  = solid triangles, mutant  $\Delta hutD$  = open squares, mutant  $\Delta hutCD$  = open triangles, (A) = M9 salts plus glucose and ammonia, (B) = M9 salts plus urocanate and (C) = M9 salts plus histidine,  $\Delta$ =deleted gene, image adopted from [150]

## 7.1 Histidine utilization (Hut) pathway

In bacteria, histidine is degraded in nature through one of four pathways but in bacteria this is normally by (Hut pathway I) or a five-step enzymatic pathway (Hut pathway II) with glutamate as the mutual end product. The two pathways have the first three steps in common from histidine to urocanate, urocanate to imidazolone propionate (IPA), and IPA to formiminoglutamate (FIGLU), catalyzed by histidase (HutH), urocanase (HutU) and imidazolone propionase (HutI), respectively (Figure 7.3).

In Hut pathway I breakdown of formimino glutamate (FIGLU) to glutamate and formamide is mediated by a single enzyme but in Hut pathway II FIGLU is converted to N-formyl L-glutamate (NFGLU) and ammonia by FIGLU iminohydrolase (HutF) encoded by *hutF* and then NFGLU is converted to glutamate and formate by formyl glutamase (HutG, also known as NFGase) encoded by *hutG*, an additional ammonium is produced in Hut pathway II. Excess ammonium is a stress for bacteria as it sends a nitrogen-replete signal causing the cessation of many catabolic enzymes including the *hut* enzymes. It has been established that some enteric bacteria, such as *Salmonella typhimurium*, *Klebsiella aerogenes* and the Gram positive bacterium *Bacillus subtilis* use the four step histidine degradation pathway (Hut I), whereas *Pseudomonas putida*, *Pseudomonas auroginosa* and *Streptomyces coelicolor* use the five-step (Hut II) pathway {150}.





**Figure 7.3 (A) The histidine degradation pathway (II), the five-step histidine degradation pathway (Hut II) of *Pseudomonas* is shown with the enzyme involved in each reaction, image adopted from Karp et al. [151]**

## 7.2 Putative role of *HutD*

A number of hypotheses have been put forward in relation to the function of HutD. Different studies have been carried out in order to prove the hypothesis, yet none have been fully conclusive.

- 1) If HutD (PAD) acts via protein-protein interaction with HutC as predicated by their genetic organization, the double deletion mutant  $\Delta hutCD$  would display the same phenotype as  $\Delta hutC$  mutant, but the  $\Delta hutCD$  mutant expressed an intermediate phenotype compared to  $\Delta hutC$  and  $\Delta hutD$  thereby indicating that HutC is not required for HutD function and functions independently of HutC [150].

- 2) It has been proposed that HutD may be involved in the regulation of *hut* operon. Growth and fitness assays by using laboratory media suggest that HutD may function to govern the upper level of *hut* transcription by controlling the intracellular concentration of the *hut* inducer (urocanate), and thus may play a regulatory role in expression of hut enzymes {150}.
- 3) HutD may bind to the weak promoter region of *hutG* and thereby repressing expression of *hut* enzymes {150}.

Based on above hypothesis and studies it is not unreasonable to expect that HutD may bind another metabolite which is not in the Hut pathway to regulate ammonium and glutamate production. This provides a difficult case for pharmacophore searching as there is no active site residues, no enzyme activity that can be assayed and potentially a significant number of ligands that could interact with it. We aim to identify a subset of chemicals which we could assess for binding and thereby identify a ligand and potentially a function to HutD.

### 7.3 Aims and objectives

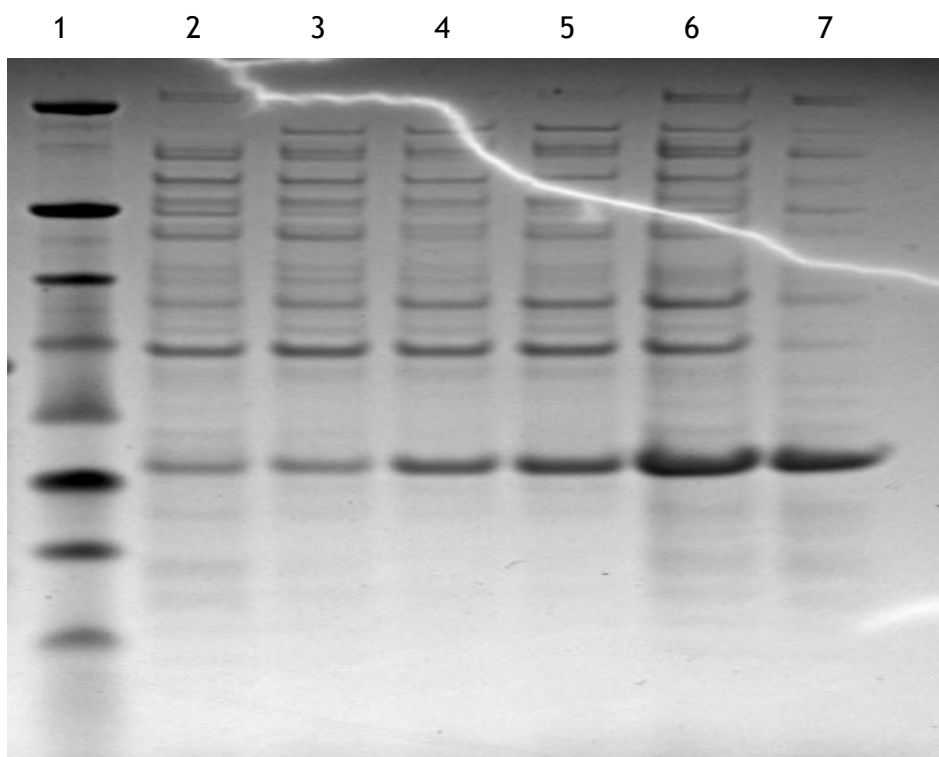
The pharmacophore searching and the ligand binding studies may provide insights into the molecular mechanism of HutD action. The objectives of this study were:

- 1) To express and purify HutD from *P. auroginosa*
- 2) To perform pharmacophore searching on the protein and obtain a small subset of potential ligands
- 3) To obtain the chemicals to test binding between HutD and potential ligands

### 7.4 Over-expression and purification of HutD

Transformations were carried out to introduce the pTBL2:PAD1 plasmid into an expression strain BL21 DE3 cells. A large numbers of colonies were obtained as a result on tetracycline plates. Over expression of HutD was carried out using LB, auto induction and M9 media. A SDS-PAGE analysis showed that about 50% of the protein was in the insoluble fraction when using LB media (Figure 7.4). In order

to increase the expression levels, different strategies were adopted, which are mentioned below. For all the SDS-PAGE analysis the NuPAGE Novex (Invitrogen®) Bis-Tris 4-12% gel was run in 4-12% Bis-Tris MES buffer.



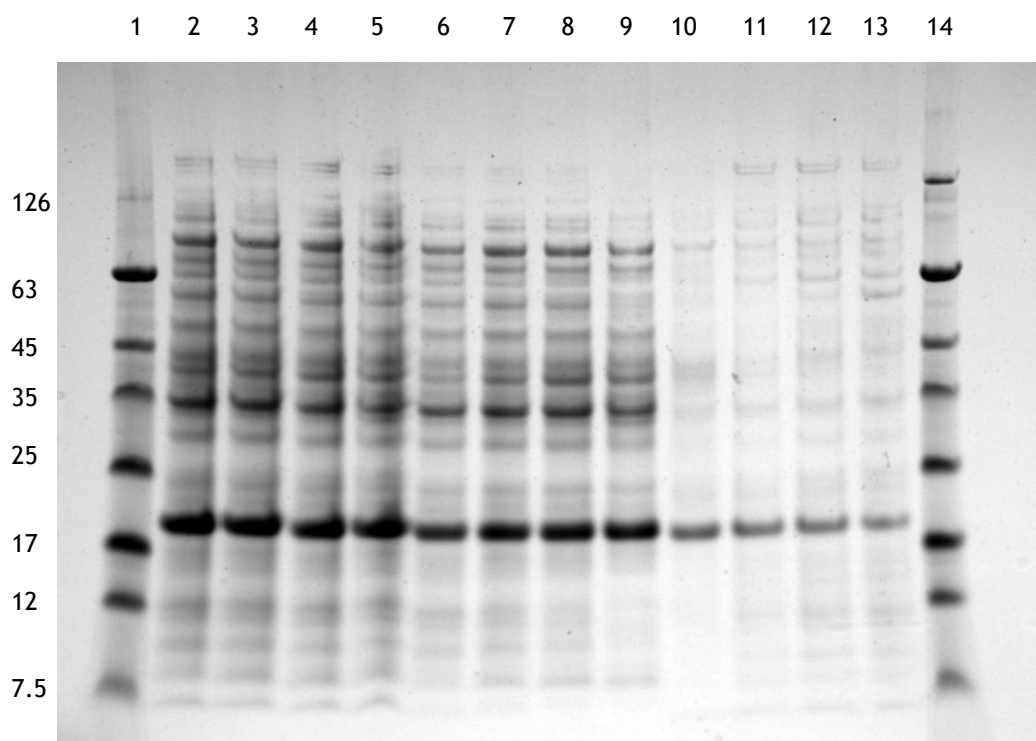
**Figure 7.4 SDS-PAGE analysis for test of expression of HutD in LB media:** 1= molecular weight marker in Kda (Bio labs, NEW ENGLAND cat# P7708S), 2=pre-induced soluble sample, 3= 1hour induced soluble sample, 4= 2hours induced soluble sample, 5=3hour induced soluble sample, 6= 4hour induced total sample, 7= 4 hour induced insoluble sample.

#### ***7.4.1 Using Auto induction media***

In order to increase the expression levels of the protein, auto induction media was used, and cultures were grown at 37°C for 24 hours. The use of auto induction media slightly increased the expression levels. The protein yield from 1L cultures was 12mg in total while from LB media it was 10mg in total.

#### ***7.4.2 Using M9 media***

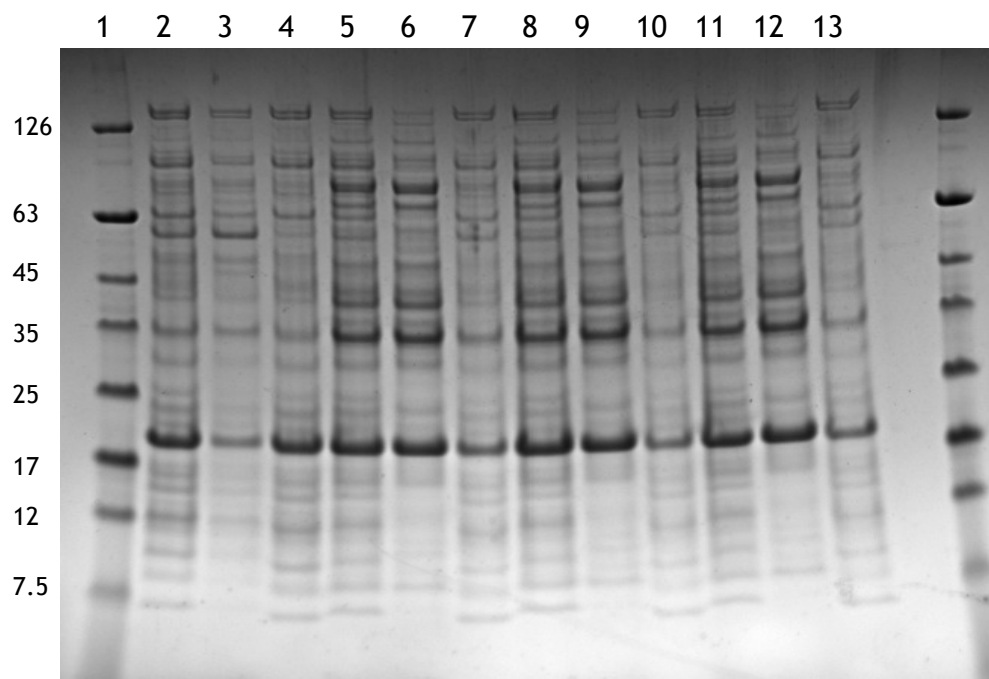
M9 media was used for growing cultures. The use of M9 media helped in increasing the expression levels (Figure 7.5). The protein yield form 1L culture was raised to 15mg in total.



**Figure 7.5 SDS-PAGE analysis for test of expression of HutD in M9 media:** 1= molecular weight marker in Kda (Bio labs, NEW ENGLAND cat# P7708S), 2=total pre-induced sample, 3= 2hour induced total sample, 4= 4hours induced total sample, 5=overnight induced total sample, 6= soluble pre-induced sample, 7= 2hour induced soluble sample, 8= 4hour induced soluble sample, 9=overnight induced soluble sample, 10=pre-induced insoluble sample, 11=2hour induced insoluble sample, 12=4hour induced insoluble sample, 13=overnight induced insoluble sample, 14=molecular weight marker in Kda

### ***7.4.3 Changing temperature and IPTG concentration of the cultures***

At higher temperatures the bacterial cultures grow faster but protein cannot fold properly and lead to the formation of misfolded protein which gets packaged into inclusion bodies and thus most of the protein remains in the insoluble fraction. On the other hand the decrease in temperature results in slow expression but low final cell density, but allows the protein to fold properly. To get around this problem large cultures (in M9 media) were grown initially at 37°C and as the O.D600 reached to 0.6, the cultures were put in slushy ice box for 5 minutes and then induced with 0.1, 0.4 and 1.0mM IPTG and left for overnight induction at 15°C. The change in temperature decreased the amount of protein in the insoluble fraction with increase in amount of protein in the soluble fraction. The varying amount of IPTG did not affect the expression level (Figure 7.6).



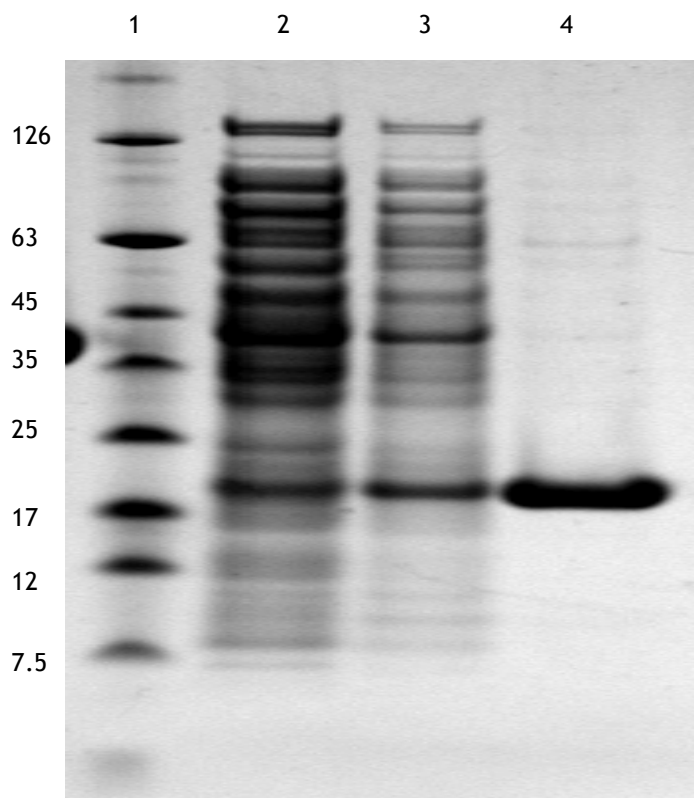
**Figure 7.6 SDS-PAGE analysis for test of expression of HutD in M9 media:** 1= molecular weight marker in Kda (Bio labs, NEW ENGLAND cat# P7708S), 2= pre-induced total sample, 3=pre- induced soluble sample, 4= pre-induced insoluble sample, 5,8,11=overnight induced total sample, 6,9,12= overnight induced soluble sample, 7,10,13= overnight induced insoluble sample, 14=molecular weight marker in Kda, 5-7= samples with 0.1mM IPTG, 8-10= samples with 0.4mM IPTG, 11-13= samples with 1.0mM IPTG

#### ***7.4.4 The effect of temperature on HutD growth***

The protein expression in host cells is usually influenced by various growth factors such as temperature, inducer concentration, induction time and duration [152]. The effect of growth temperature on HutD solubility was examined and the results showed significant effect of growing temperature on protein solubility: when HutD was induced at 15°C instead of 37°C, a large proportion of HutD was soluble. The protein yield reached to 25mg from 1L culture as a result of inducing the protein at low temperature. The effect can be explained by the possibility that high temperatures could enhance interaction of hydrophobic regions of protein folding intermediates and promoting the formation of inclusion bodies whereas lower growth temperatures may reduce the expression rate of unfolded proteins and thereby decreasing the kinetics of protein intermediate interaction, and thus limiting the aggregate formation and enhancing the protein solubility [153] .

### 7.4.5 Harvesting and Purification of *HutD*

The cells were harvested and further nickel purification was carried out as described in section 2.13. Different fractions obtained from Ni-column were loaded on a SDS PAGE gel (Figure 7.7). The target protein was found in the elution fraction (Elution buffer: 300mM Imidazole 50mM Tris, 300mM NaCl pH: 7.5). To get rid of high imidazole content the elution fraction was subjected to dialysis for 4 hours in 20mM Na-phosphate, pH:7.5 dialysis buffer and later on concentrated down at 3000xg by using Vivaspin® and then passed through the gel filtration column as a final purification step.



**Figure 7.7 SDS-PAGE analysis for *HutD* after Ni-purification:** 1= molecular weight marker in Kda (Bio labs, NEW ENGLAND cat# P7708S), 2=flow through, 3= 75mM Imidazole wash, 4= 300mM Imidazole elute

### 7.4.6 Size exclusion chromatography (Gel Filtration)

The concentrated sample was then subjected to size exclusion chromatography as described in section 2.13.2. During Gel filtration it was observed that the protein gave a single sharp peak when at a concentration of 1mg/ml (Figure 7.8). While at a higher concentration (15mg/ml) an additional shoulder peak also appeared attached to the main peak (Figure 7.9). This indicated that

protein tends to oligomerize at a higher concentration and has both the monomer and dimer (higher content) forms while at lower concentration only monomer form exists. This was confirmed by running a SDS PAGE gel on eluted fractions in which the bands for fractions corresponding to the shoulder peak appeared at the same location as for the main peak (Figure 7.10). The evidence from size-exclusion chromatography suggests that HutD forms a dimer in the elution buffer at higher concentrations, is in agreement with the dimerization of HutD reported in *P. fluorescens* SBW25 {154}. The pure fractions obtained after gel filtration were pooled and concentrated down to the desired concentration by centrifuging at 4000xg in a viva spin. The absorption co-efficient for the protein was calculated by submitting the amino acid sequence of the protein in ExPASy ProtParam tool (<http://web.expasy.org/protparam/>), which was 1.76 for 1mg/mL. The final concentration of protein was calculated by measuring the absorbance at 280nm by using JASCO-550 UV spectrophotometer.

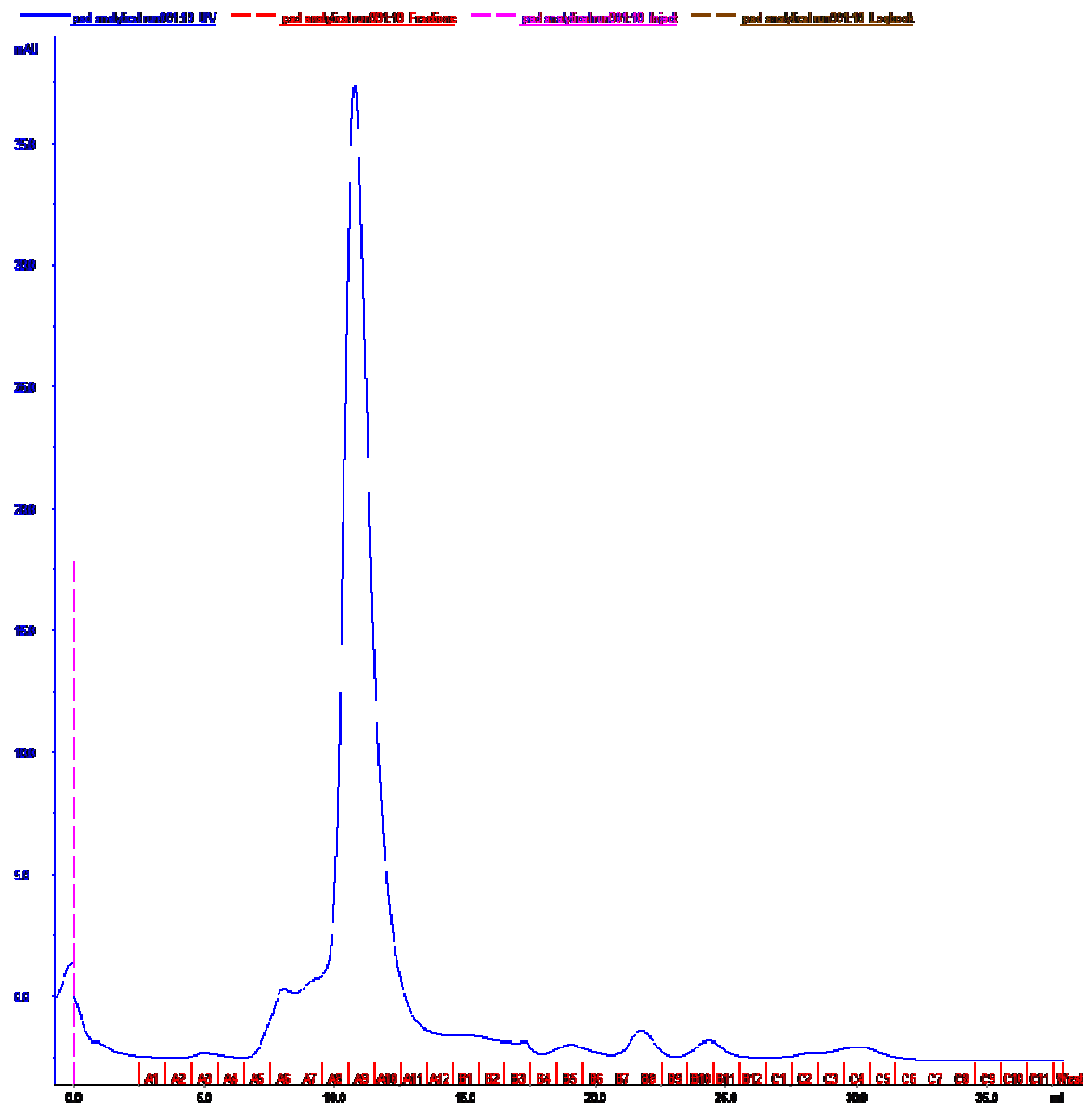


Figure 7.8 Chromatogram for HutD after size exclusion chromatography with individual fractions eluted on molecular size basis (at 1mg/ml the protein elutes after 12ml, suggestive of the monomer form)



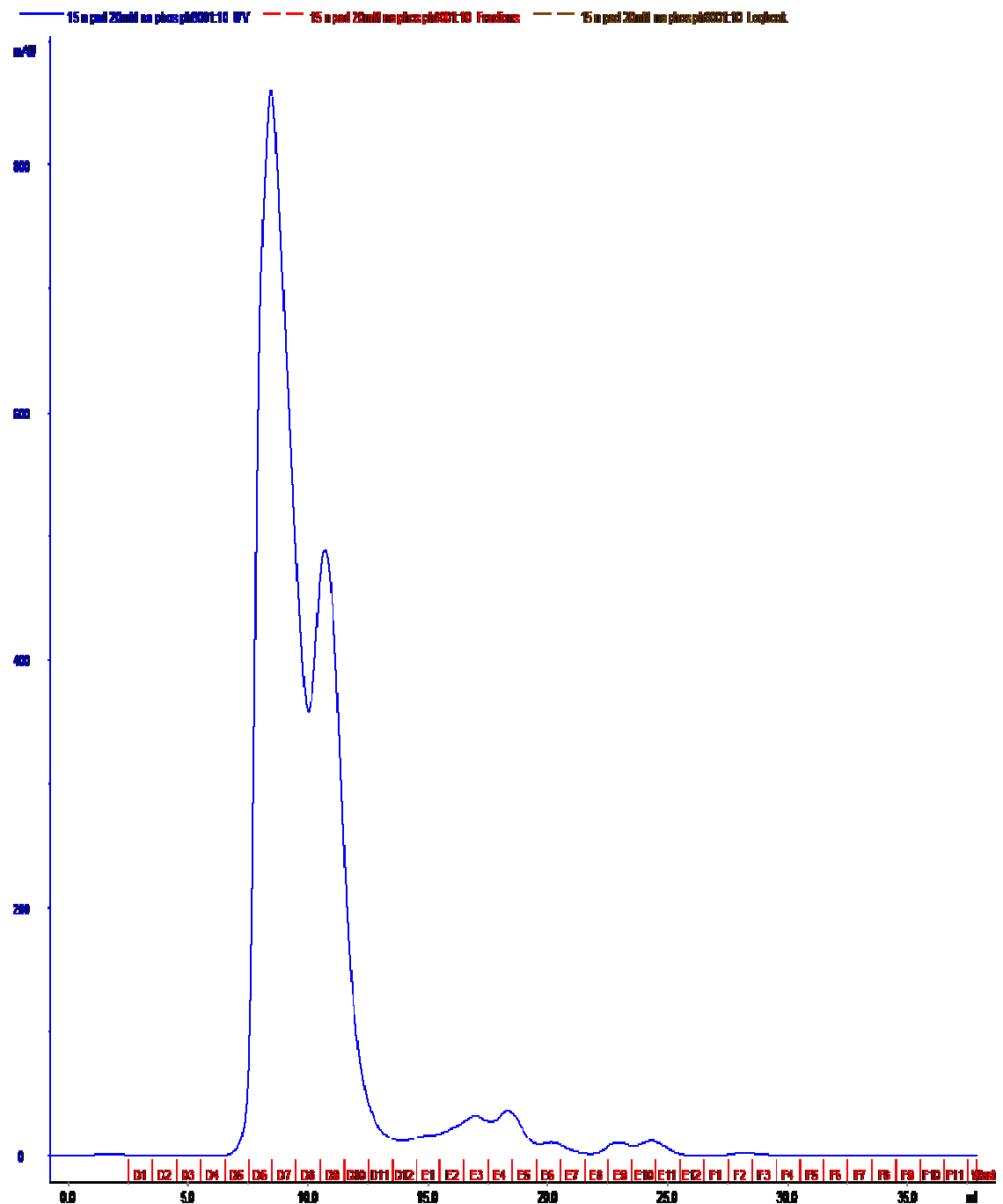
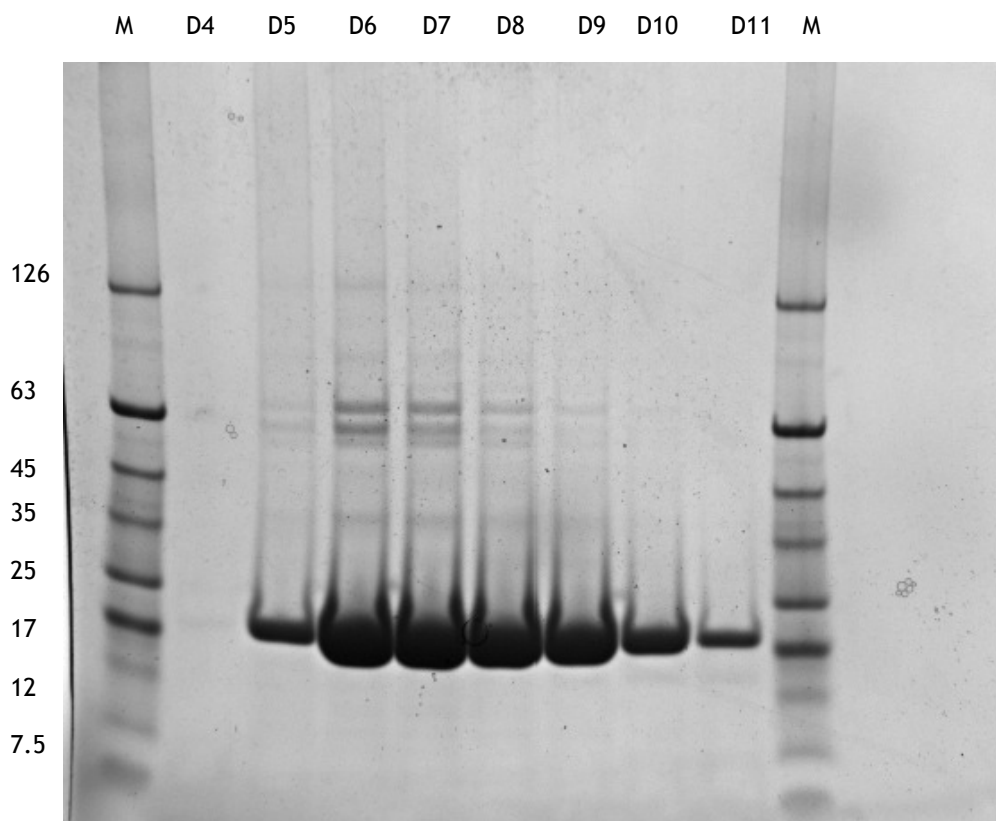


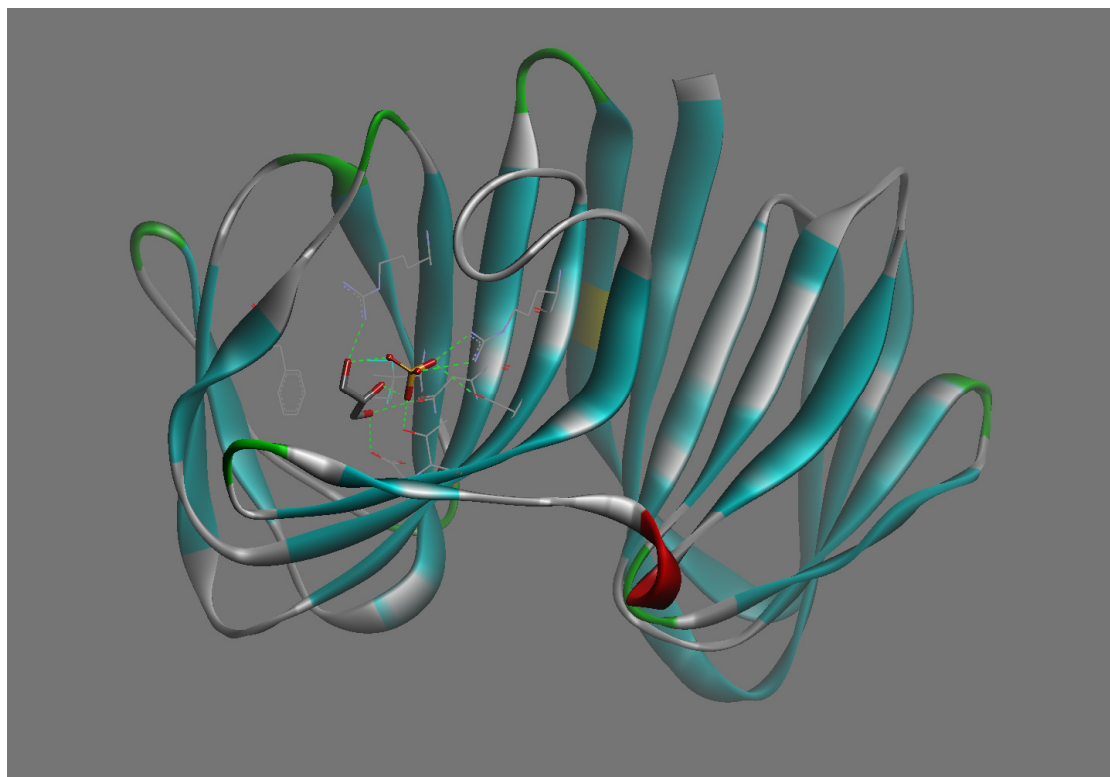
Figure 7.9 Chromatogram for HutD after size exclusion chromatography with individual fractions eluted on molecular size basis (at 15mg/ml the protein elutes at 9ml (dimer form)with a shoulder peak at 12ml,suggestive of the monomer form)



**Figure 7.10 SDS-PAGE analysis for HutD after gel filtration: M= molecular weight marker in Kda (Bio labs, NEW ENGLAND cat# P7708S), D5-D11 = pure elution fractions containing HutD**

## 7.5 Pharmacophore searching for HutD

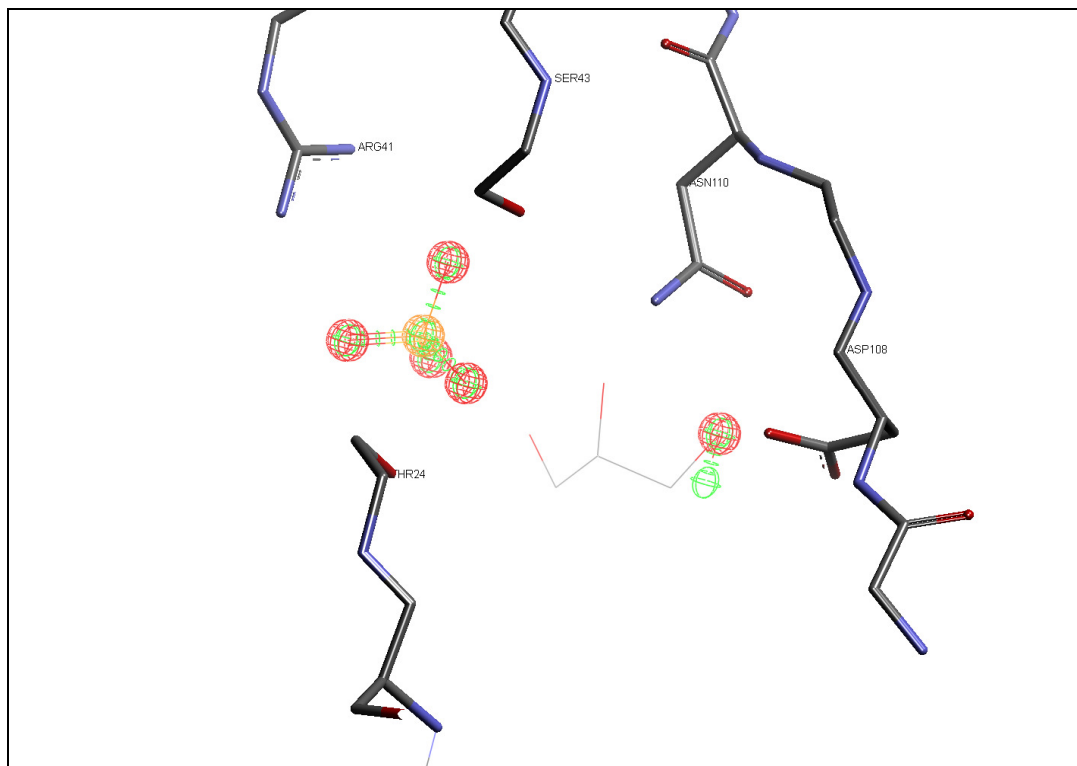
Pharmacophores had been run previously with the hydrogen bond vector method with little or no success. With improvements in the pharmacophore methodology described in this thesis, future pharmacophores for HutD were generated by using the query atom method as described in section 3.5.3. All the preliminary searches were carried out in catalyst by using the “fast flexible search” option prior to the actual pharmacophore search through the ChEBi database. If the search gave hits then the actual search was carried out against naturalism database by using Accelrys catalyst®. Figure 7.11 shows the solid ribbon diagram of HutD along with the Sulfate group and glycerol molecule in the putative binding site, forming important H-bonds with the active site residues.



**Figure 7.11** Solid ribbon diagram for HutD along with Sulfate and glycerol molecule in the putative binding site, forming Hydrogen bonds with the active site residues of the protein.

### **7.5.1 Pharma HutD 1**

As a  $\text{SO}_4^{2-}$  ion and a glycerol molecule were found in D chain of the protein in one of our unpublished HutD crystal structure (Lokhamp Bernhard PhD thesis 2003, University of Glasgow). The first pharmacophore for HutD was generated for D chain of the protein and the query ligand was generated by using it as a template (Figure 7.11). The hydrogen count on oxygens of the query ligand was set to 1 and the radius of the uncertainty location sphere for query atoms was set to  $0.4\text{\AA}$ . The pharmacophore was then subjected to search against the naturalism database. We presumed that a sugar phosphate would be a suitable ligand.



**Figure 7.12** Query ligand in the active site of HutD with surrounding key amino acid residues the fragment type, query ligand generated by using  $\text{SO}_4$  and hydroxyl group of the glycerol as template (all exclusion spheres, some amino acid residues and water molecules have been removed for clarity viewing)

All the hits contained a phosphate group aligned with the query atoms (generated via sulphate ion). Maximum hits had their hydroxyl group satisfying the query atom of the pharmacophore (generated via hydroxyl group of glycerol). Majority of the hits included phospho sugars like beta-D-fructose 6-phosphate, D-xylulose 5-phosphate, D-ribulose 5-phosphate, erythrulose 1-phosphate. some hits had phosphate groups attached to glycerol like glycerol 1-phosphate, 3-phospho-D-glyceroyl dihydrogen phosphate, 2,3-bisphosphoglyceric acid, 3-phosphoglyceric acid, 3-phospho-D-glyceric acid while few phospho amines like serine phospho ethanolamine, glycerol-3-phospho ethanolamine were among the hits as well.

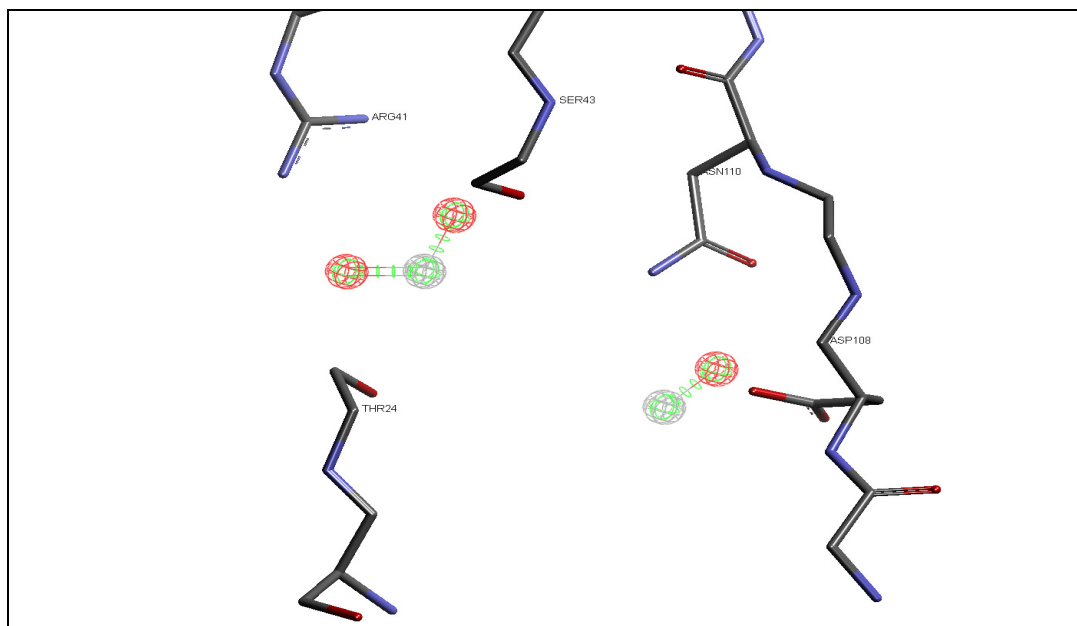
**Number of hits obtained = 71**

**Time taken = 40 minutes**

### **7.5.2 Pharma HutD 2**

The hits obtained through first pharmacophore had short contacts with the surrounding amino acids, which ruled out the possibility of H-bonding and some hits were having hydrophobic clash with certain residues and thus unlikely to be

the ligands. In order to avoid this, the pharmacophore was modified by increasing the uncertainty location sphere around query atoms to 0.6Å. With the hypothesis that the ligand may have H-bond interactions with the conserved arginine (Arg41) of the protein, the query ligand sulphate fragment was modified to carboxylate group, with its 2 oxygens pointing towards the Arg41 of the protein. Further in the crystal structure the binding site has glycerol, forming good H-bonds with ASP108 and ASN110 therefore carbon was added to OH group of glycerol in the pharmacophore (Figure 7.12). The database search just took 1 minute.



**Figure 7.13 Modified pharmacophore with oxygens of the query ligand carboxylate group pointing towards ARG41 and SER43 at H-bond distance, carbon atom added to the hydroxyl group query ligand towards the ASP108 (all exclusion spheres, certain amino acid residues and water molecules have been removed for clarity viewing)**

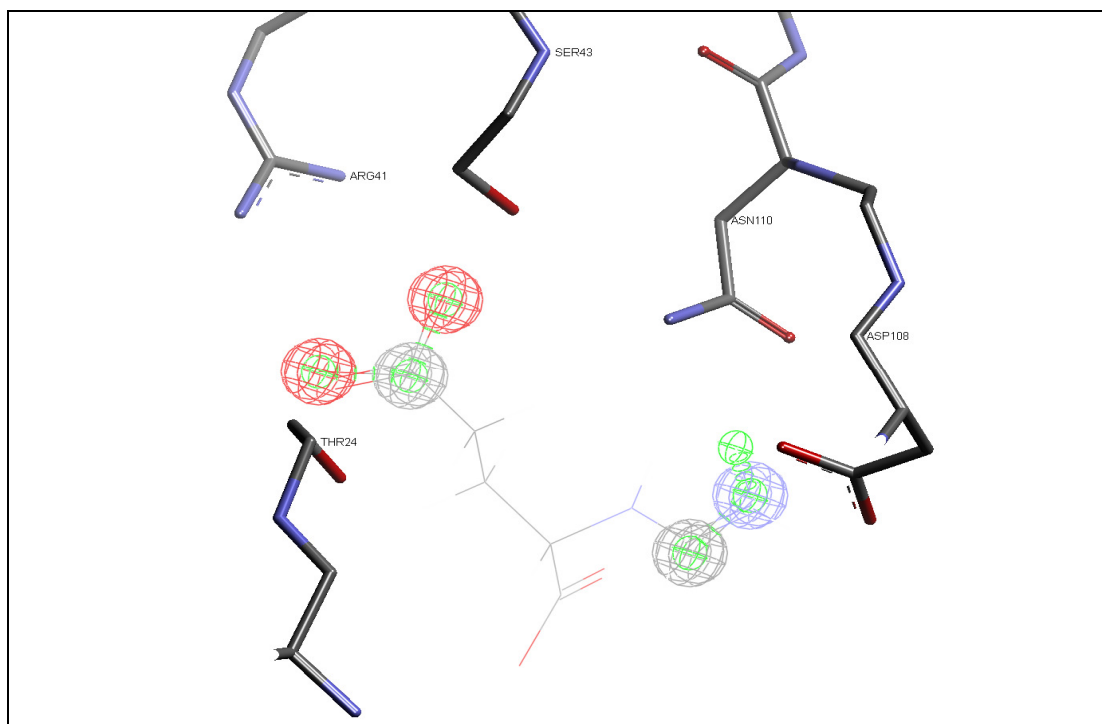
This time the hits were not in clash with the active site residues and formed good H-bonds with residues like ASP108, ASN110 towards the glycerol query end and with ARG41 towards the carboxylate query end. The hits included sugars like D-galacturonate, D-arabinonate and 5-dehydro-2-deoxy-D-gluconate. The hits also included some dicarboxylate compounds like 5-hydroxypentanoate, 3, 4-dicarboxy-3-hydroxybutanoate and glutarate.

**Number of hits obtained = 12**

**Time taken = 1 minute**

### 7.5.3 Pharma HutD 3

In order to check the possibility of different types of ligands, the oxygen atom attached to carbon (towards the ASP108 residue) was changed to Nitrogen and the hydrogen count on Nitrogen was set to free. The bond between carbon and nitrogen was optimized to single/double (Figure 7.13). The final pharmacophore was then searched through the naturalism database.



**Figure 7.14** Final pharmacophore with oxygen atom of the query fragment towards ASP108 replaced with Nitrogen, FIGLU is shown as line model satisfying the query and making H-bond interactions with ASP108 and ARG41 (all exclusion spheres, certain amino acid residues and water molecules have been hidden for clarity viewing)

The hits from the final pharmacophore included mostly amino acid and their derivatives like D-arginine, L-arginine, arginine, N-formimidoyl-L-glutamate (FIGLU) and some guanidine compounds like 3-guanidinopropanoic acid, 4-guanidinobutanoic acid and 5-guanidino-2-oxopentanoate. The database search interestingly gave Formimino glutamate (FIGLU) among the hits (important metabolite in the Hut pathway). The formimino group of FIGLU formed H-bond with the carboxyl group of ASP108. In addition the carboxylate of FIGLU which was not searched for did not make any clashes but rather was in the proximity of residues with which it could make hydrogen bonds. Based on pharmacophore model it was considered that FIGLU could be the most likely ligand to bind to this protein.

Number of hits obtained = 9

Time taken = 1 minute

## 7.6 Ligand selection and synthesis

The selection and synthesis of ligands was based on the hits obtained from pharmacophore searching and metabolites involved in Hut II pathway. FIGLU an important metabolite of the Hut pathway was the most convincing hit that came out from pharmacophore search. NFGLU and glutamate are the penultimate and last metabolites respectively in the Hut II pathway, their structural similarity to the predicted hit FIGLU made them obvious choices for comparison. Urocanate was included for binding studies as it has been predicted {154} to bind to the protein through ITC studies but not by pharmacophore searching. Based on growth studies {150} and pharmacophore searching histidine has been suggested not to be a ligand therefore was included as a negative control.

### 7.6.1 Synthesis of N-Formimino L-Glutamate (FIGLU)

N-Formimino L-Glutamate (FIGLU) was synthesised by the procedure used for the preparation of N-formimino L-aspartic acid by Hayaishi {155}. The method involved the condensation of formamidine and L-Glutamate in the presence of  $\text{Ag}_2\text{CO}_3$  and formamide. To prepare FIGLU, 1.06g (20mmol) of formamidine hydrochloride, 2.0g (11mmol) of  $\text{Ag}_2\text{CO}_3$ , 1.12g (10mmol) of L-glutamic acid, and 5mL of formamide were vigorously stirred for 36 hours at room temperature in 50mL round bottom flask, stoppered with a  $\text{CaCl}_2$  tube. The mixture was then subjected to rotary evaporation for 30 minutes to remove most of the dissolved  $\text{CO}_2$ , treated with 100mL of 0.06N HCl and then filtered. The filtered solution was then subjected to rotary evaporation. The final fraction obtained after rotary evaporation was treated with ethanol to remove excess of formamide (as formamide is soluble in ethanol). The insoluble fraction obtained after treatment with ethanol was subjected to freeze drying to get pure form of FIGLU as described in biochemical preparations {156}.

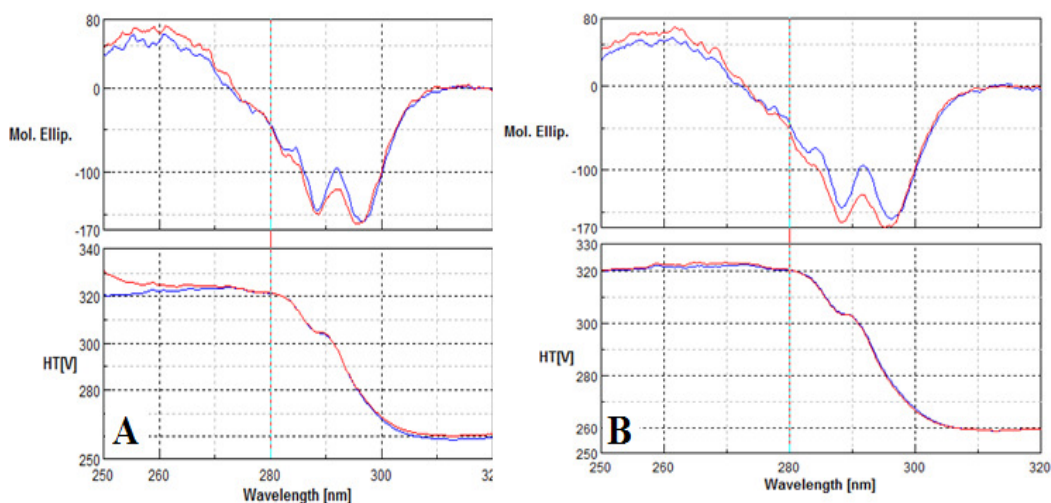
$^1\text{H}$  NMR and  $^{13}\text{C}$  spectra were recorded on a Bruker DPX-400 spectrometer with chemical shift values in ppm relative to TMS ( $\delta_{\text{H}}$  0.00 and  $\delta_{\text{C}}$  0.0) as standard with  $\text{D}_2\text{O}$  as a solvent at room temperature. Spectroscopic data was in

accordance with the literature.  $\delta_H$  (400 MHz,  $D_2O$ ) 2.00 (2H, q,  $J$  7.9 Hz, 3- $CH_2$ ), 2.26 (2H, t,  $J$  4.1 Hz, 4- $H_2$ ), 4.1 (1H, t,  $J$  6.2 Hz, 2-H), 7.78 (1H, s, 5-COOH), 7.80 (1H, s, 1-COOH);  $\delta_C$  (100MHz,  $D_2O$ ) 27.53, 32.70, 57.43, 153.67, 175.05 and 180.83.

## 7.7 CD results for HutD

To see the effect of potential ligands on the secondary structure of the protein, particularly on the aromatic regions of the protein the selected potential ligands were added to the protein and their near and far UV spectra were recorded. All the CD spectra were recorded at a protein concentration of 1.82mg/mL (for Near UV) and 0.92 mg/mL (for Far UV). Cell path length was 0.2cm and 0.01cm for Near and Far UV analysis respectively. The CD experiments were carried out in buffer solution of 20mM Na-phosphate, 50mM NaCl pH: 7.8 at 20°C.

Significant changes were observed in the Near UV CD spectra as a result of addition of FIGLU and NFGLU (Figure 7.14). Though the shifts were much bigger in case of NFGLU, the changing pattern in the near UV region (particularly in 290-300nm range) for NFGLU and FIGLU was pretty similar.



**Figure 7.15** Near UV (250-320nm) CD spectra for (A) HutD (blue) and HutD with 5mM FIGLU (red), (B) HutD (blue) and HutD with 5mM NFGLU (red)



The bigger shifts in the near UV region due to the addition of NFGLU suggest some changes in the secondary structure of the protein. Further the similarity in the Near UV CD spectra might indicate that these ligands are causing the same effect in or around the binding site of the protein. This indicates that the ligand may have caused some change in the arrangement of the aromatic amino acid residues as a result of direct or indirect interaction. No significant changes were observed in the far UV region of the CD spectra due to addition of NFGLU and FIGLU (Figure 7.15)

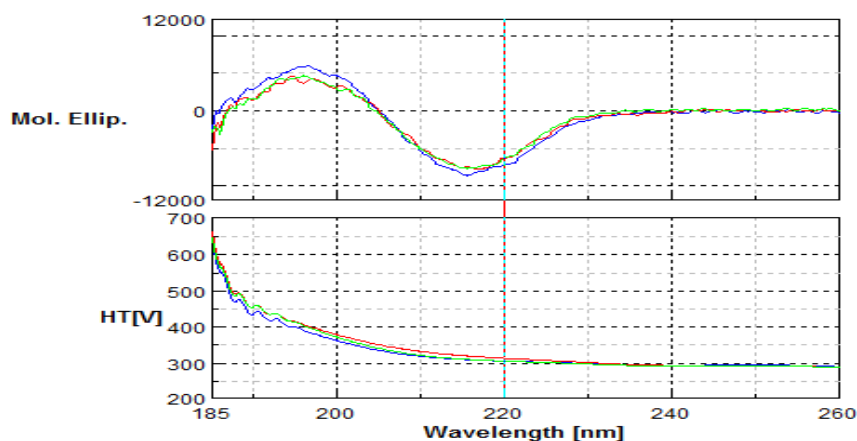


Figure 7.16 Far UV (190-260nm) CD spectra of HutD (blue) with 2.5mM FIGLU (red) and 2.5mM NFGLU (green).

The spectral changes brought by the addition of L-glutamate and L-Histidine were not observable. The pattern observed for both these ligands was same in the Near UV and Far UV region (Figure 7.16-17). The overall composition of  $\alpha$ -helices and  $\beta$ -strands nearly remained the same in case of these ligands.

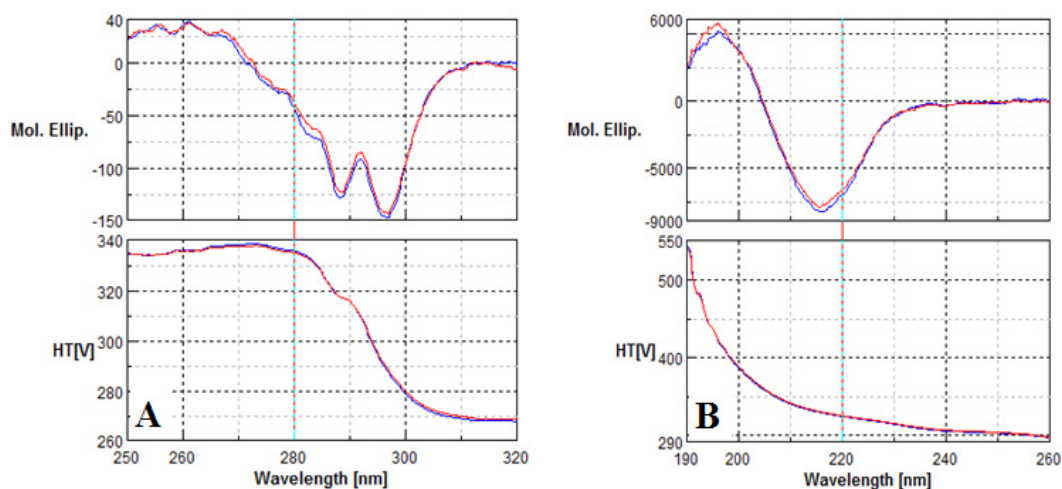


Figure 7.17 (A) Near UV (250-320nm) and (B) Far UV (190-260nm) CD spectra for HutD (blue) and HutD with 310 μM L-glutamate (red)

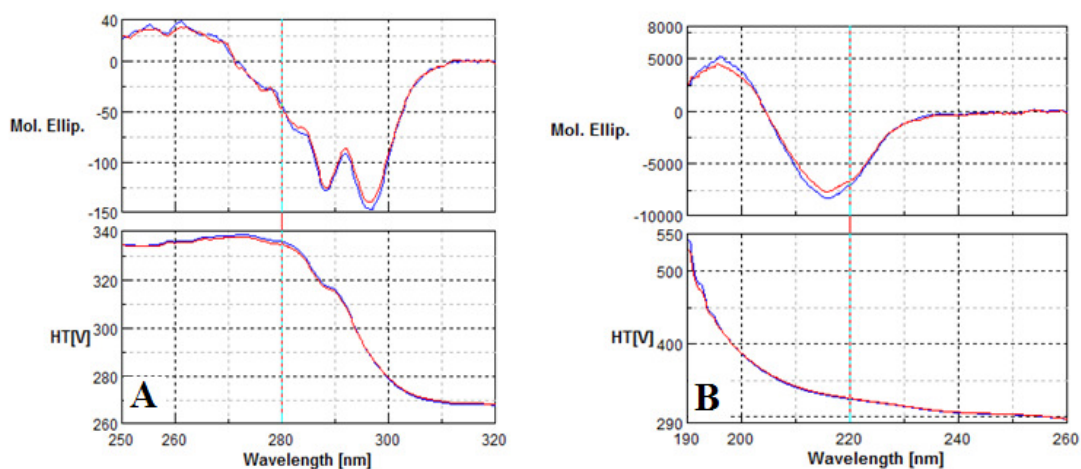


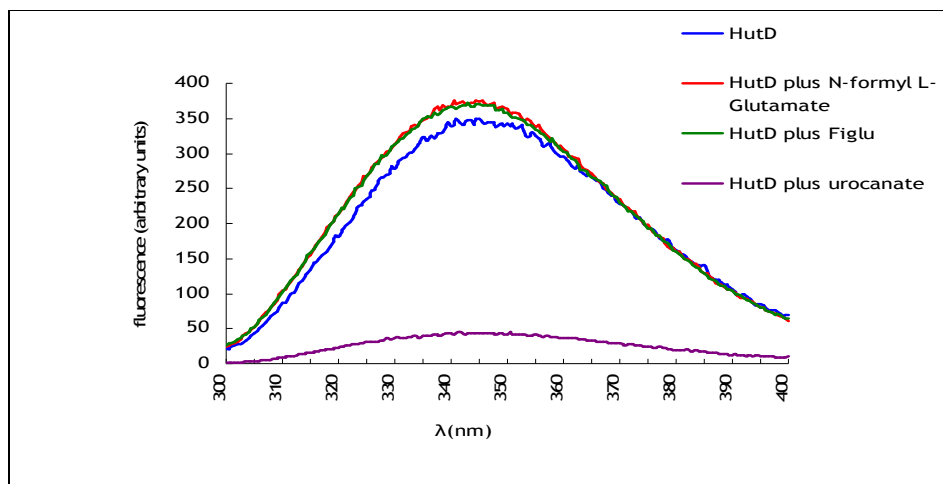
Figure 7.18 (A) Near UV (250-320nm) and (B) Far UV (190-260nm) CD spectra for HutD (blue) and HutD with 310 μM L-Histidine (red)

In case of urocanate the experimental studies were hampered due to the fact that urocanate absorbs very strongly in the Near UV and Far UV region. In order to test and confirm that urocanate absorbs the UV very strongly, the fluorescence experiments were carried out.

## 7.8 Fluorescence spectroscopy on HutD

When a molecule is transformed from ground state to excited state through electromagnetic radiation, the energy of absorbed photon is equal to the energy difference between two levels, the molecule can return to its ground state by emitting a photon of energy difference to that of the excited state, this emission is known as fluorescence. The electromagnetic radiation is typically absorbed by a tryptophan in a protein {157}. As there are 6 tryptophans in HutD structure, fluorescence experiments were performed to monitor conformational changes in protein structure as a result of interaction with ligands. The concentration of HutD in all samples was 0.1mg/mL and concentration of all ligands was 0.27mM. The buffer used was 20mM Na-phosphate pH: 7.8. All the fluorescence experiments were performed on Perkin Elmer LS508 spectrofluorimeter.

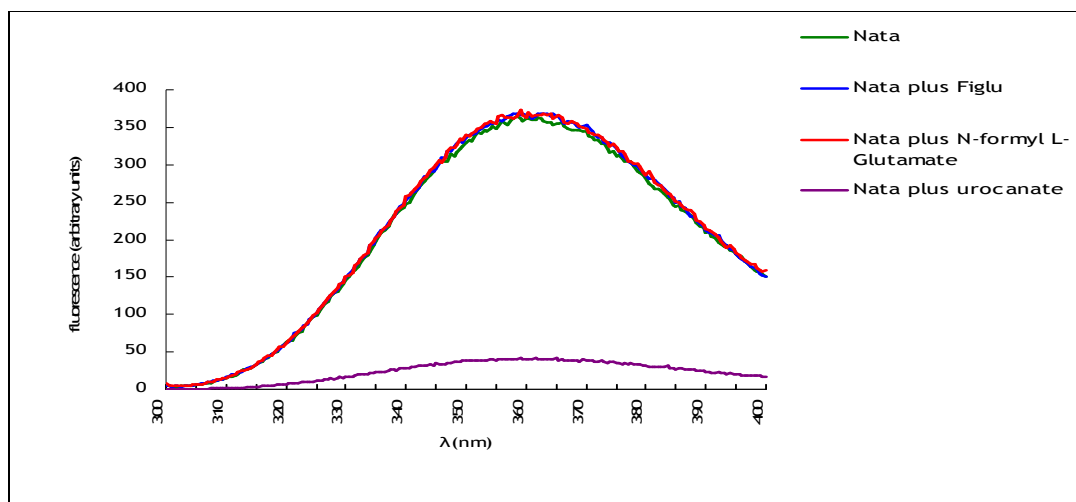
Only a little increase in intensity of fluorescence was observed when FIGLU and NFIGLU were added to the protein (Figure 7.18). In case of urocanate a big decrease in fluorescence was observed which could be due to the quenching effect of urocanate.



**Figure 7.19** Overlaid fluorescence spectra for HutD with potential ligands (the quenching effect is prominent due to the addition of urocanate to HutD)

In order to confirm the quenching effect, a control experiment was performed. N-acetyl-tryptophan-amide (NATA) solution (0.025mg/mL) was prepared in the same buffer and the ligands were added one by one to see the effect. Here also when urocanate was added to NATA, the huge quenching effect was observed

(Figure 7.19). No spectra was obtained showing any binding of ligand causing major structural change in the protein.



**Figure 7.20** Overlaid fluorescence spectra for N-acetyl-tryptophan-amide (Nata) with different ligands (the quenching effect is prominent due to the addition of urocanate to Nata)

## 7.9 NMR results for HutD

NMR had been shown to be successful for testing the ligand binding for TdcF (Chapter: 6), therefore HSQC spectra were carried out by using the  $^{15}\text{N}$ -labelled protein. However, the spectra were not clean as might have been predicted from the crystal structures. It was therefore necessary to optimize the conditions for getting better and cleaner spectra. Initially the spectra were recorded at a range of temperature with varying buffer composition (salt and pH). Throughout the optimization it was observed that some of the peaks in the spectra were smeared, which could be because of the dimerization of the protein, and thus causing a change in the structure at the dimer interface. The dimerization of the protein has previously been indicated during the purification of the protein at gel filtration step, but not in earlier studies (Bernhard Lokhamp PhD thesis 2003, University of Glasgow). The protein samples for NMR both with and without the ligand were prepared as described in section 6.6.1. After optimization of the conditions all the HSQC spectra for the protein were recorded in 20mM Na-phosphate, pH:7.2 buffer at 43°C. The ligands were prepared in the same buffer solution as of the protein and further the pH of the

ligand solution was adjusted prior to their addition to the protein. The concentration of protein was 300 $\mu$ M in all the experiments.

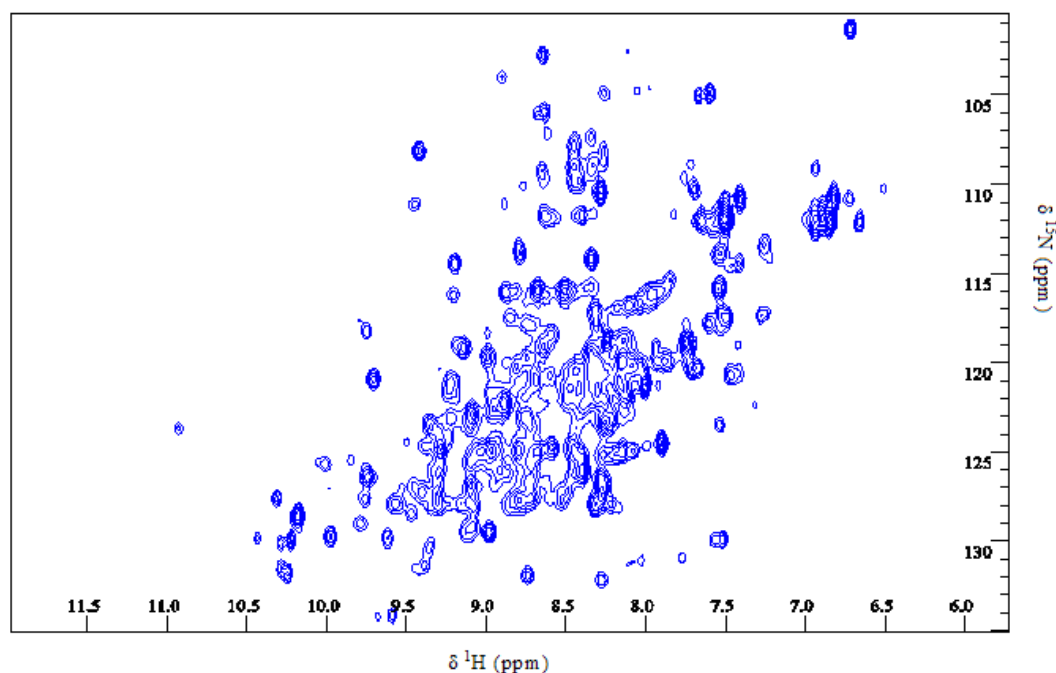
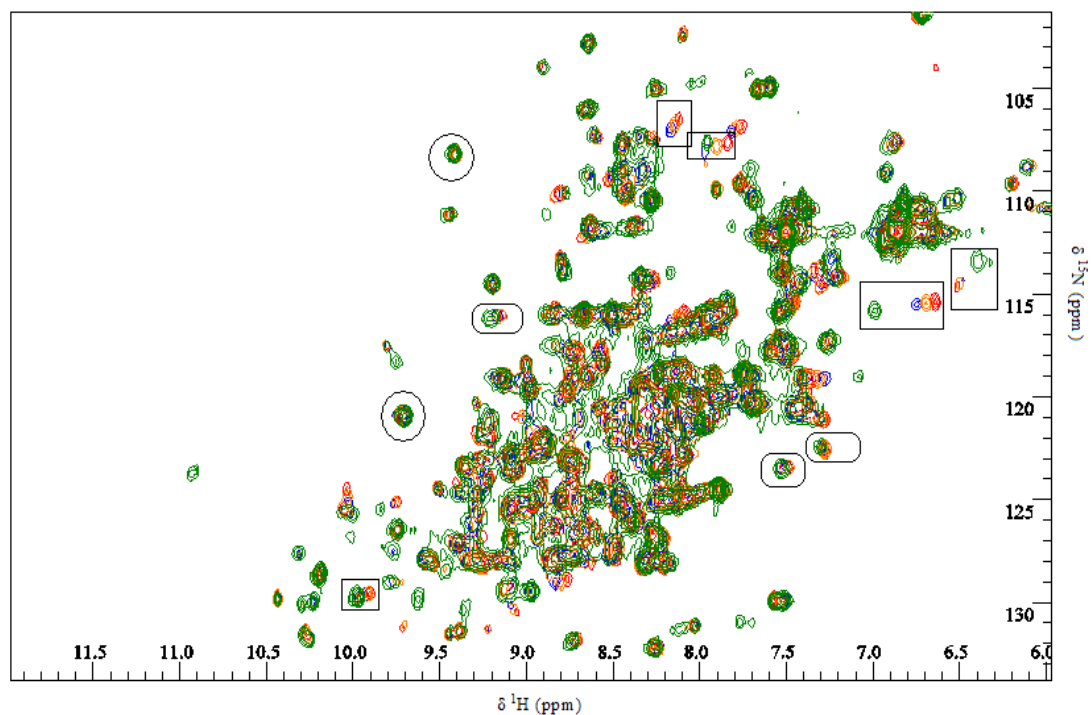


Figure 7.21 NMR-HSQC spectra for  $^{15}$ N-labelled HutD without any ligand

### 7.9.1 HutD with FIGLU

The overlaid spectra of protein with successive additions of ligand showed some marked changes. A total of 15 peaks changed their position. It was observed that certain peaks slightly changed their positions while few peaks travelled larger distances (Figure 7.21). Certain peaks moved a lot with increasing concentration of the ligand, as represented by open square boxes around the peaks. The range of ligand concentration was used between 0-15mM. At 15mM ligand concentration certain peaks disappeared and some new peaks appeared which might suggest some conformational change in the structure of the protein as a result of ligand binding. In native protein spectrum some smeared peaks (corresponding to two peaks) appeared which could be attributed to two conformations of protein as a result of dimer formation. In the subsequent spectra with the addition of FIGLU certain portion of the smeared peaks appeared stronger with increase in intensity which suggested preference of one conformation over the other as a result of ligand binding. The protein

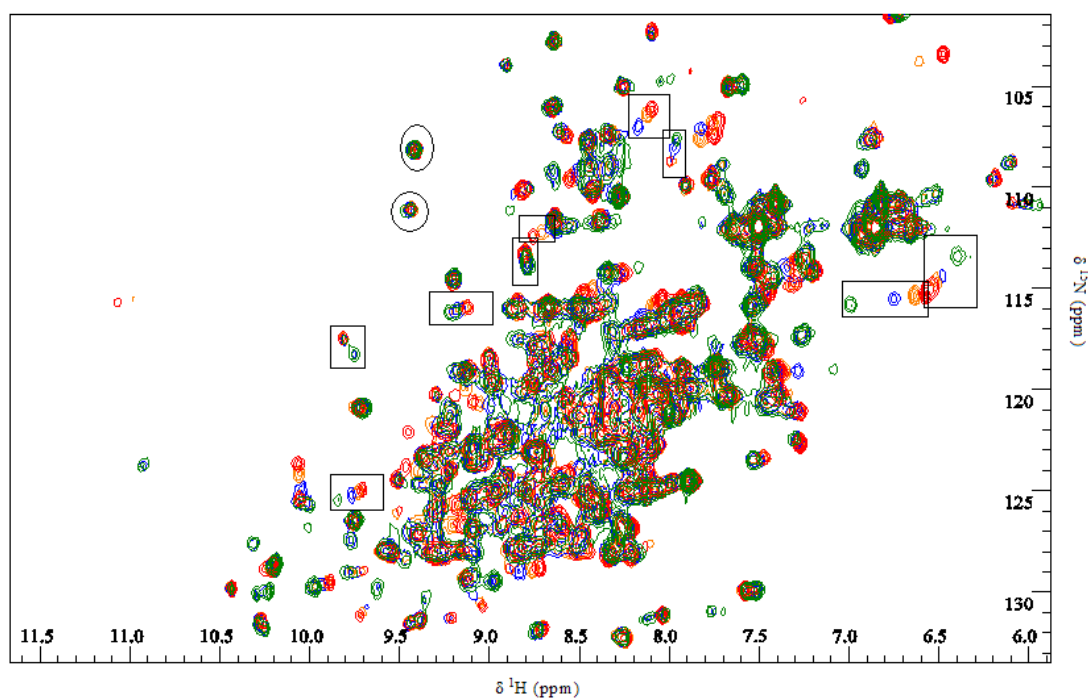
seemed to be saturated at 10mM ligand concentration with the ligand as at 15mM ligand concentration the peaks did not shifted any further



**Figure 7.22** Overlaid HSQC spectra of HutD with FIGLU, empty square boxes= peaks bigger shift due to the addition of the ligand, empty circles= peaks with no change due to the addition of ligand, rounded rectangles= peaks with smaller shifts (green spectra is for protein alone, blue orange and red spectra are at 3,5 and 10mM ligand Concentration)

### 7.9.2 *HutD* with *NFGLU*

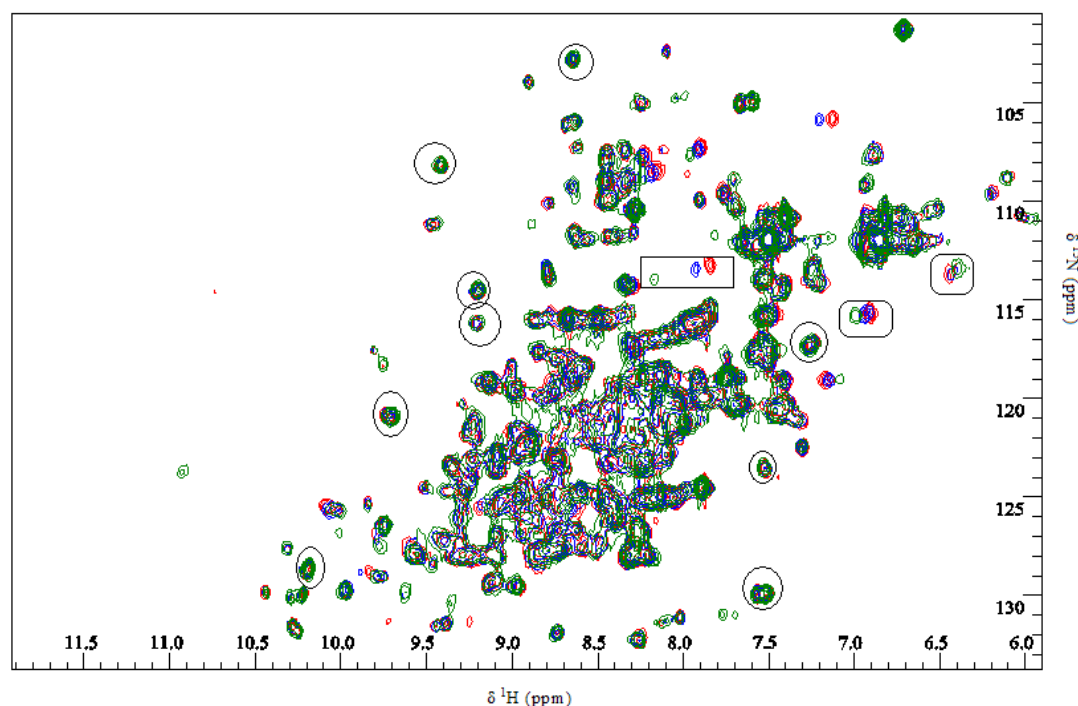
The comparison of spectra between the protein alone and protein with NFGLU showed marked changes with change in position of 10 peaks. It was observed that certain peaks slightly changed their positions and few peaks disappeared due to the addition of the ligand (Figure 7.22). Some new peaks also appeared at the start of ligand addition and further kept on moving with successive addition of ligand. For some peaks the pattern of movement was similar as to FIGLU. The protein was saturated at 10mM ligand concentration as further addition of ligand did not cause any change in the peak positions.



**Figure 7.23** Overlaid HSQC spectra of *HutD* with *NFGLU*, empty square boxes= peaks bigger shift due to the addition of the ligand, empty circles= peaks with no change due to the addition of ligand, rounded rectangles= peaks with smaller shifts (green spectra is for protein alone, blue orange and red spectra are at 1, 3 and 10mM ligand concentration)

### 7.9.3 *HutD* with Urocanate

By seeing the overlaid spectra after the addition of urocanate, it was observed that 5 peaks shifted their position, among which only one peak travelled a big distance and the rest shifted slightly. By comparing the spectral changes brought by Urocanate it appeared that the specific peaks which changed their position due to Urocanate also shifted in case of NFGLU and FIGLU but the shift distance travelled due to the former was fairly small in comparison to latter one. The protein appeared to be saturated at 10mM ligand concentration as further addition of ligand did not cause any significant changes in the position of the peaks (Figure 7.23).



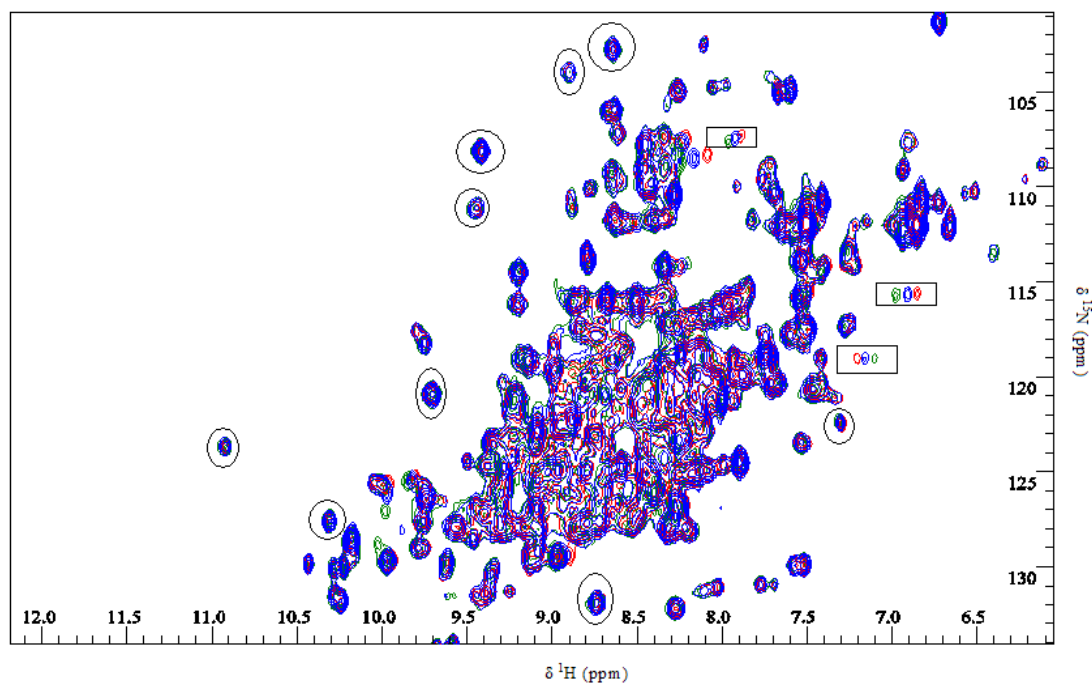
**Figure 7.24** Overlaid HSQC spectra of *HutD* with Urocanate, empty square boxes= peaks bigger shift due to the addition of the ligand, empty circles= peaks with no change due to the addition of ligand, rounded rectangles= peaks with smaller shifts (green spectra is for protein alone, blue and red spectra are at 5 and 10mM ligand concentration)

By comparing the shift changes in peaks due to the addition of FIGLU, NFGLU and Urocanate were more or less of the same pattern. The changing effect observed due to NFGLU was more obvious (peaks travelling bigger distances) than by FIGLU and Urocanate for the same peaks.



### 7.9.4 *HutD* with *L*-Glutamate

The changes observed as a result of addition of *L*-Glutamate to the protein were fewer than with FIGLU and NFGLU. Only 4 peaks appeared to shift due to ligand addition while rest remained static (Figure 7.24). The specific peaks which changed their position due to *L*-glutamate also appeared to be shifting in case of NFGLU and FIGLU but the shift distance travelled due to the former was fairly small in comparison to latter one. The ligand was used in the concentration range of 0–20mM.



**Figure 7.25** Overlaid HSQC spectra of *HutD* with *L*-glutamate, empty square boxes= peaks bigger shift due to the addition of the ligand, empty circles= peaks with no change due to the addition of ligand, (green spectra is for protein alone, blue and red spectra are at 10 and 20mM ligand concentration)

### 7.9.5 *HutD* with *L*-Histidine

In case of *L*-histidine, the spectra showed no change in the peak positions except slight shift in only one peak (Figure 7.25). Interestingly the addition of all the other ligands did not change the position of this peak. A logical explanation for this could be the charge state of Histidine, as Histidine has a pKa value of 6.5 and small increase in pH can cause it to be deprotonated. The pH of the buffer solution was 7.2 which means that Histidine can cause protonation of some protein residues. This suggests that *L*-histidine either does not bind to the protein or binds very weakly even at a concentration of 20mM.

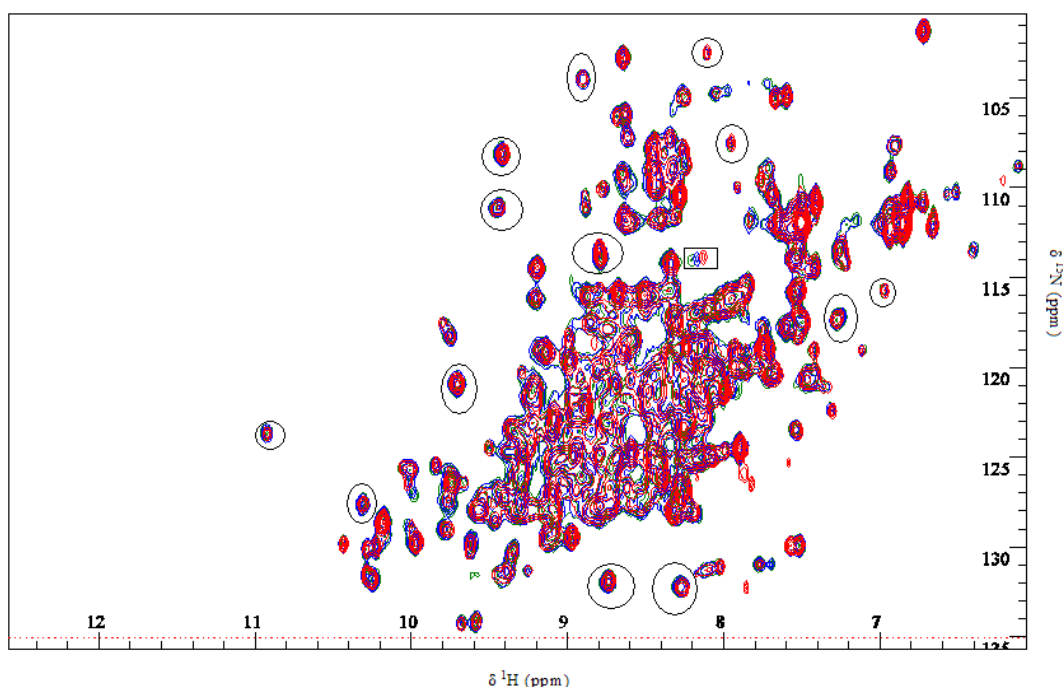


Figure 7.26 Overlaid HSQC spectra of *HutD* with *L*-Histidine, empty square boxes= peaks bigger shift due to the addition of the ligand, empty circles= peaks with no change due to the addition of ligand (green spectra is for protein alone, blue and red spectra are at 10 and 20mM ligand concentration)

### 7.9.6 Determination of $K_d$ values

The  $K_d$  values for different ligands were calculated (Table 7.1) by using the same fitting function as described in section 6.6.2. In terms of  $K_d$  values NFGU appeared to be a tighter binder than FIGLU. For Urocanate and *L*-glutamate  $K_d$  values were in the mM range, which shows weak binding. The  $K_d$  value for *L*-Histidine could not be determined due to insufficient data points to fit the curve

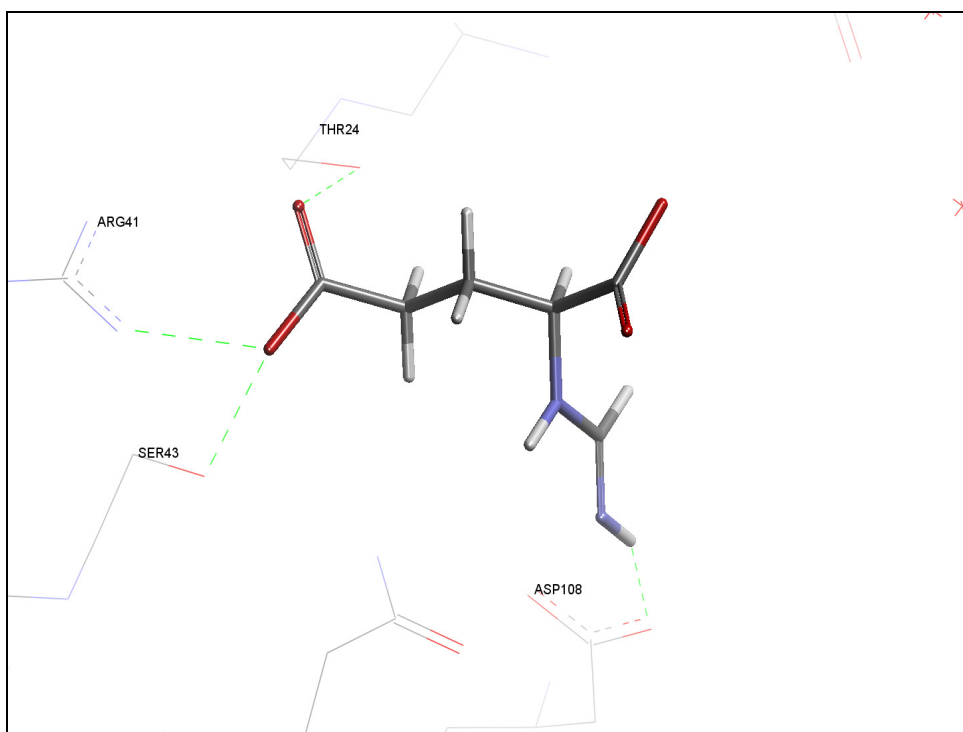
however the HSQC spectra for histidine indicates that it does not bind to the protein.

S.#	Ligand	Ligand concentration range (mM)	K <sub>d</sub> for HutD
1	FIGLU	0-15	500 $\mu$ M $\pm$ 0.58
2	NFIGLU	0-15	92 $\mu$ M $\pm$ 0.25
3	L-glutamate	0-20	>30mM
4	Urocanate	0-10	>1.5mM
5	L-Histidine	0-20	N.D

**Table 7.1** The range of different ligands added to HutD along with their calculated K<sub>d</sub> values

## 7.10 Conclusions and future work

1. The pharmacophore search extracted FIGLU as a potential ligand form the database, and predicted a discrete binding mode for this ligand (Figure 7.26).



**Figure 7.27** Graphical representation of FIGLU (stick model) in the binding site of HutD, green broken lines show FIGLU forming H-bonds with THR24, ARG41, SER43 and ASP108

2. The availability of pseudo ligands in the form of crystallization chemicals was important as it allowed the use of the query atom method. However, it did not predict NFGLU as a ligand because we had no evidence that Asp 108 could be protonated and so to act as a proton donor for the carbonyl oxygen of NFGLU.
3. From CD and NMR experiments it can be deduced that HutD does not significantly interact with histidine, urocanate or glutamate. There is strong evidence of binding to NFGLU and FIGLU with NFGLU having a  $K_d$  in the physiological range for ligand.
4. From the NMR results it appears that NFGLU causes more shifts in the spectra than FIGLU which need to be characterized more to confirm which is the preferred ligand.
5. The outcome of the NMR and CD experiments was in agreement for ligands like NFGLU, FIGLU and L-Histidine. The results helped to differentiate between true binders and non-binders. Both the techniques suggest L-Histidine to be a non-binder.
6. This work shows that HutD binding to the ligands in descending order is

**NFGLU > FIGLU > Urocanate**

This further supports the hypothesis that HutD may function as a binding protein, to play a regulatory role in the Hut pathway.

7. The favorable ability of HutD to crystallize means that either soaking of native crystals or co-crystallisation with ligands should permit a detailed look at the binding of both FIGLU and NFGLU and whether they elicit the same conformation in the protein. Obtaining a clear idea of the ligand that binds to HutD should allow a better understanding of the function of this protein and how it performs a regulatory role in histidine utilization pathway.

## 8. *Ralstonia eutropha* N-formylglutamate amido-hydrolase like protein (PARI)

As part of an ongoing project, studying the structure and mechanism of *Pseudomonas aeruginosa* amidohydrolase (PAA), a N-formyl glutamate amido hydrolase (NFGase) from the histidine utilization pathway. A crystal structure of a related protein (PDB code: 2Q7S) from *Ralstonia eutropha*, annotated as NFGase (PARI) and deposited in protein databank under the JCSG project. The sequence alignment (Figure 8.1) carried out between PARI and NFGases from different microorganisms demonstrates significant sequence similarity over the whole protein sequence, suggesting a common fold and function. The x-ray structure of PARI has been determined at a resolution of 2.0Å, and is annotated as NFGase, solely on the basis of sequence similarity and over presumed fold similarity to the actual NFGases. The protein comprises of 290 amino acids with a molecular weight of 32.6 Kda and exists in the monomer form.

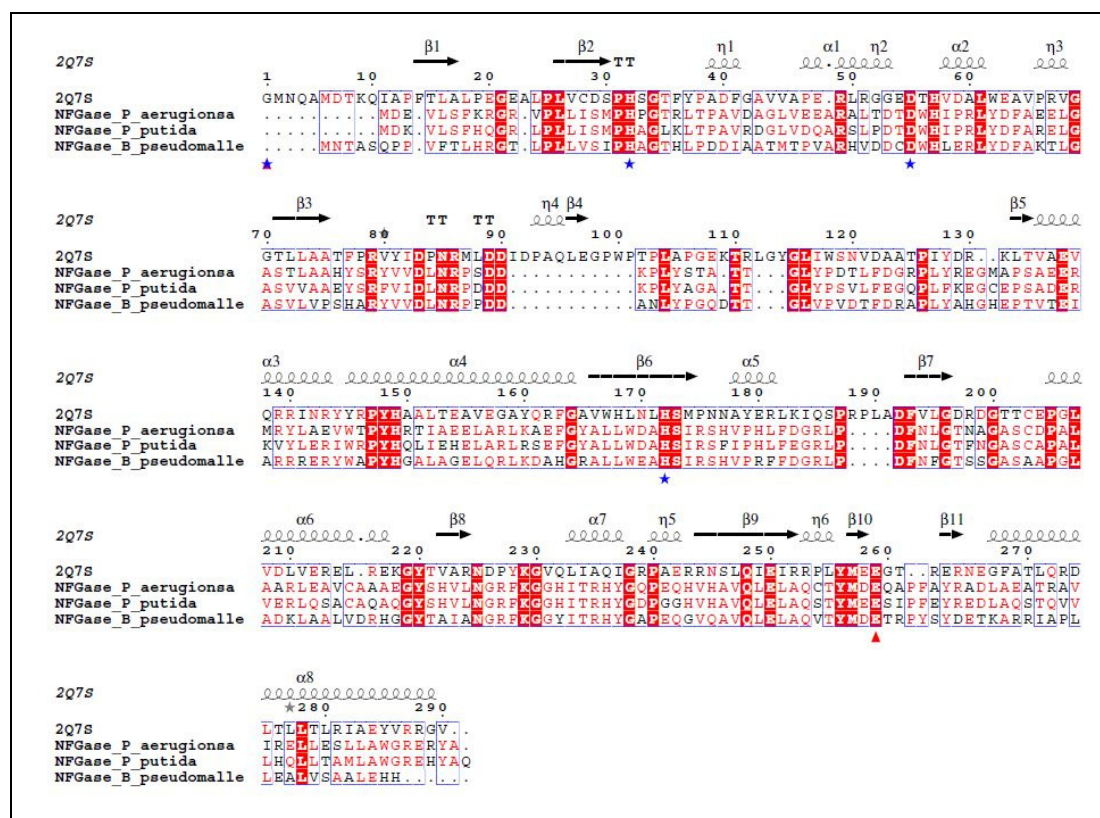


Figure 8.1 Amino acid sequence comparison of PARI with other NFGases, residues highlighted with red bars indicate the position and location of identical residues in the structure, the sequences were aligned using MultAlin [158]

As co-crystallization experiments were unsuccessful with PAA while enzyme inhibition assays demonstrated nano molar inhibition with certain inhibitors (as described in section 9.4). It was decided to study *R. eutropha* enzyme (PARI) which could be more suitable for structural characterization. For this purpose the plasmid of PARI was purchased from the Protein Structure Initiative Material Repository (PSI-MR, ref: PMC2808882).

## 8.1 Over-expression and purification of PARI

The PARI plasmid came as a small cell culture. The cell cultures were streaked on agar plates containing kanamycin (30µg/mL). Some overnight cultures were also setup from the same cell cultures.

The plasmid was isolated from the overnight liquid cultures by using the QIAGEN® Miniprep kit. The concentration of the plasmid DNA (70ng/µL) was checked by using nano drop spectrophotometer. Transformations were carried out to introduce the plasmid into an expression strain BL21 DE3 cells. Large numbers of colonies were obtained as a result on kanamycin plates.

Over expression of PARI was carried out by using LB, auto induction and M9 media. The His-tagged protein was purified by using nickel affinity chromatography. The yield of protein from LB media was only 1–1.5mg protein from 1L culture. This was significantly less than the reported value mentioned on the web link of JCSG website. A SDS-PAGE analysis was carried out which showed that most of the protein was in the insoluble fraction, while only a small amount was present in the soluble fraction of the protein (Figure 8.2). For all the SDS-PAGE analysis the NuPAGE Novex (Invitrogen®) Bis-Tris 4-12% gel was run in 4-12% Bis-Tris (MES) buffer and for reference the molecular weight marker from Bio labs, NEW ENGLAND cat# P7708S was used. In order to increase the expression levels of the protein, different strategies were adopted which are mentioned below.

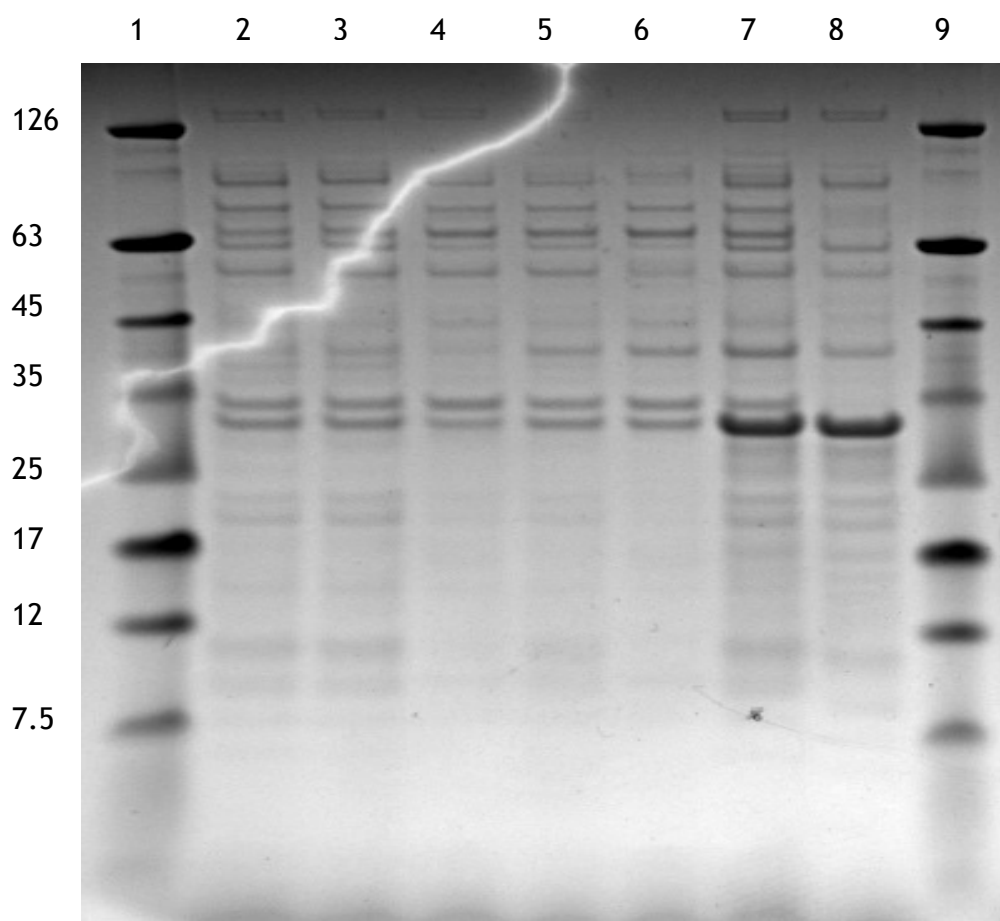
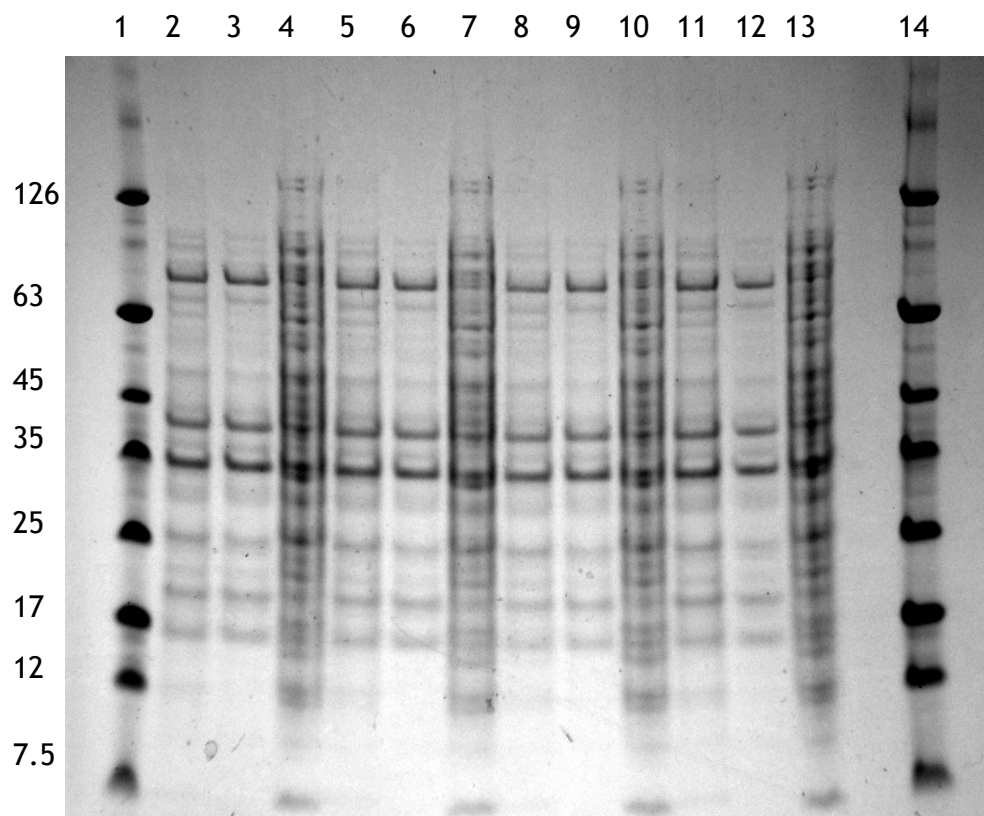


Figure 8.2 SDS-PAGE analysis for test of expression of PARI in LB media at 37°C: 1 = molecular weight marker in Kda, 2 = pre-induced soluble sample, 3= 1hour induced soluble sample, 4 = 2hours induced soluble sample, 5 = 3hour induced soluble sample, 6 = 4hour induced soluble sample, 7 = 4 hour induced total sample, 8 = 4hour induced insoluble sample, 9 = molecular weight marker in Kda (numerals in vertical column represents different molecular weights for individual bands)

### ***8.1.1 Lowering temperature of cultures in auto induction media.***

In order to increase the expression levels of the protein, cultures were grown in auto induction media at 20°C and 30°C. The lower temperature should help by reducing the rate of over expression and give more time for the bacteria to attempt to refold incorrectly folded protein. At higher temperature the mis folded protein is targeted to inclusion bodies to keep it from interfering with the functioning of the cell. The usage of auto induction media at lower temperature did not increase the expression levels significantly.



**Figure 8.3 SDS-PAGE analysis for test of expression of PARI in auto induction media after over night induction at lower temperatures: 1 = molecular weight marker in Kda, 2, 3&4 = total, soluble and insoluble samples grown at 20°C in bowelled flasks, 5, 6&7 = total, soluble and insoluble samples grown at 20°C in non-bowelled flasks, 8, 9&10 = total, soluble and insoluble samples grown at 30°C in bowelled flasks, 11, 12&13 = total, soluble and insoluble samples grown at 30°C in non-bowelled flasks, 14=molecular weight marker in Kda.**

#### **8.1.1.1 Using Rosetta™ 2 DE3 cells**

The presence of some rare codons in the plasmid of PARI, could lead to stalling of translation of the mRNA during protein synthesis and hence leading to insoluble protein. In order to prevent this transformations were carried out in Rosetta™ 2 (DE3) competent cells (Novagen material number: 71397-4, Kit batch number: D00111972, Antibiotic resistance: chloramphenicol) which provides sufficient supply of rare codons. The usage of Rosetta cells did not increase the expression levels significantly.



### 8.1.2 Using M9 media (supplemented with zinc sulphate)

The choice of growth media and using different cell types can have an effect on the over expression of proteins. This is due to a difference in the metabolites and proteins present in the cell that might help or hinder correct folding of the protein. As by using auto induction media the expression levels were too low, M9 media (minimal media) was used for growing cultures. The use of M9 media helped in increasing the expression levels and the protein yield from 1L cultures reached to 4-5mg in total. As the active site of the protein has  $\text{Zn}^{2+}$  and to rule out the possibility of  $\text{Zn}^{2+}$  as a limiting factor,  $\text{ZnSO}_4$  was added to some batches as a source of  $\text{Zn}^{2+}$  at a final concentration of  $10\mu\text{M}$ . The addition of zinc to the cultures did not significantly increase the expression levels, therefore was discontinued later on.

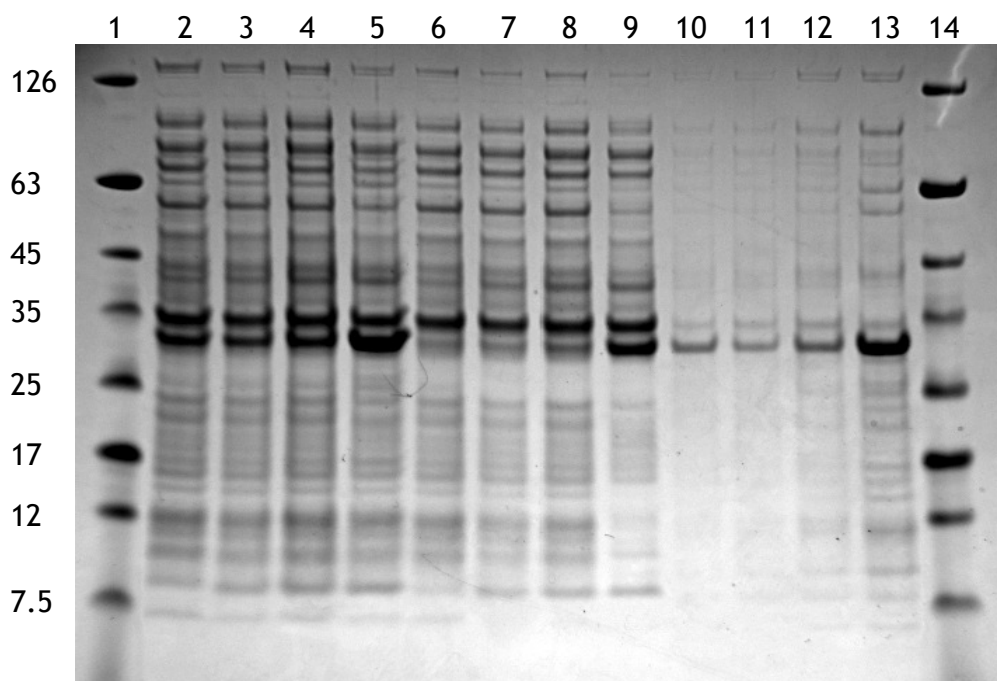


Figure 8.4 SDS-PAGE analysis for test of expression of PARI in M9 media: 1 = molecular weight marker in Kda, 2 = total pre-induced sample, 3 = 2hour induced total sample, 4 = 3hours induced total sample, 5 = overnight induced total sample, 6 = soluble pre-induced sample, 7 = 2hour induced soluble sample, 8 = 3hour induced soluble sample, 9 = overnight induced soluble sample, 10 = pre-induced insoluble sample, 11 = 2hour induced insoluble sample, 12 = 4hour induced insoluble sample, 13 = overnight induced insoluble sample, 14 = molecular weight marker in Kda.

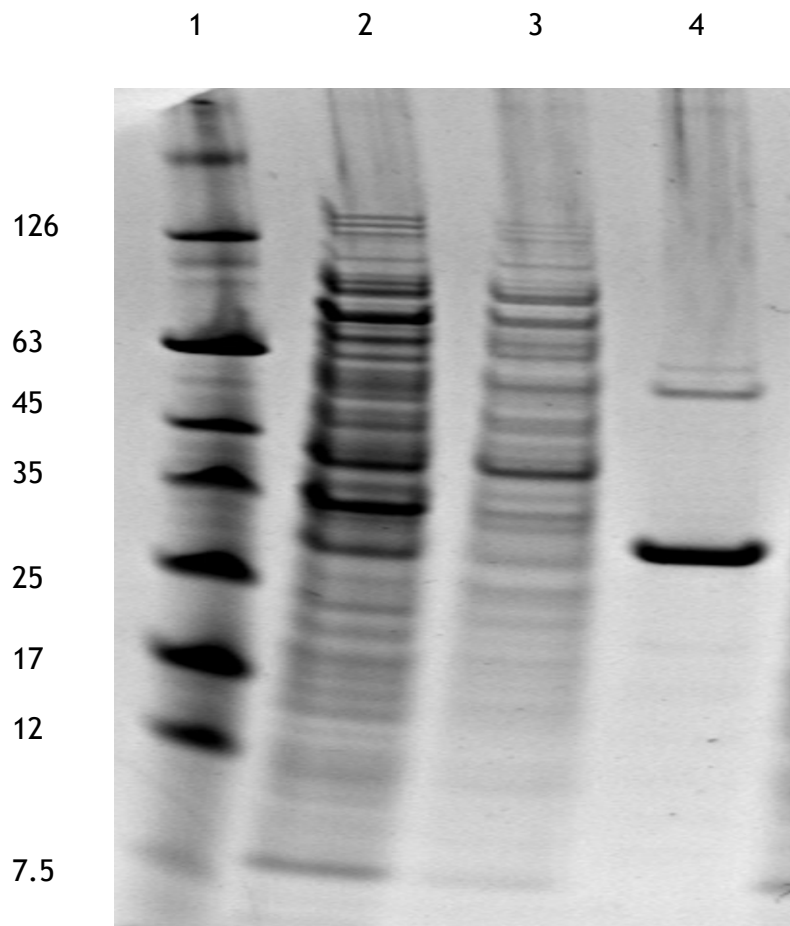
#### 8.1.2.1 Lowering temperature of M9 cultures

One of the possible reasons for low expression levels could be the misfolding of the expressed protein due to high temperature conditions of the growing

cultures, as at higher temperature the protein grows faster but cannot fold properly and some times this misfolded protein culminates in the form of inclusion bodies, leaving the protein in the insoluble fraction. To prevent these possibilities large cultures (in M9 media) were grown initially at 37°C and as the O.D600 reached 0.6, the cultures were put in slushy ice box for 5 minutes and then induced with 1mM IPTG and further left for overnight growth at 15°C. The change in temperature slightly increased the expression levels and the resulting protein yield was 4-6mg in total from 1L cultures. Poor expression levels can also be attributed to any gene which becomes toxic to the plasmid and thus damaging the T7 RNA polymerase, leading to very low yields of protein.

### ***8.1.3 Purification of PARI***

The cells were harvested and further nickel purification was carried. Different fractions obtained from Ni-column were loaded on a SDS PAGE gel (Figure 8.5). The protein was found in the elution fraction. The purified protein obtained after nickel purification had significant contamination of DNA as demonstrated by a peak at 260 nm from UV spectrum. In order to remove the high imidazole content and DNA content the elution fraction was subjected to over night dialysis in 2L of 20mM Na-phosphate buffer, pH: 7.5 at 4°C. The UV spectrum of protein sample after dialysis showed the removal of DNA with no peak at 260nm. The protein was then concentrated by using Vivaspın® at 3000xg to the desired concentration and then passed through the gel filtration column.



**Figure 8.5 SDS-PAGE analysis for PAR1 after Ni-purification: 1 = molecular weight marker in Kda, 2 = flow through, 3 = 20mM Imidazole wash, 4 = 300mM Imidazole elute**

#### **8.1.3.1 Size exclusion chromatography (Gel Filtration)**

The concentrated sample was subjected to size exclusion chromatography. The pure fractions obtained from gel filtration (Figure 8.6) were pooled and concentrated down by using the Vivaspin® at 3000xg. To check the purity of the protein the fractions obtained from gel filtration chromatography were then loaded on to SDS-PAGE (Figure 8.7). The concentration of the protein was measured by using the molar absorption value. The molar absorption value ( $A_{280} = 1.24$  for 1g/L) was obtained from the amino acid sequence of the protein by using the expasy prota param resource (<http://web.expasy.org/proparam>). It was observed that the protein concentrated more quickly at 20°C rather than at 4°C. During purification of different batches it was observed that the protein is purified better in buffer solution of 20mM Na-Phosphate, 50mM NaCl pH: 7.8.

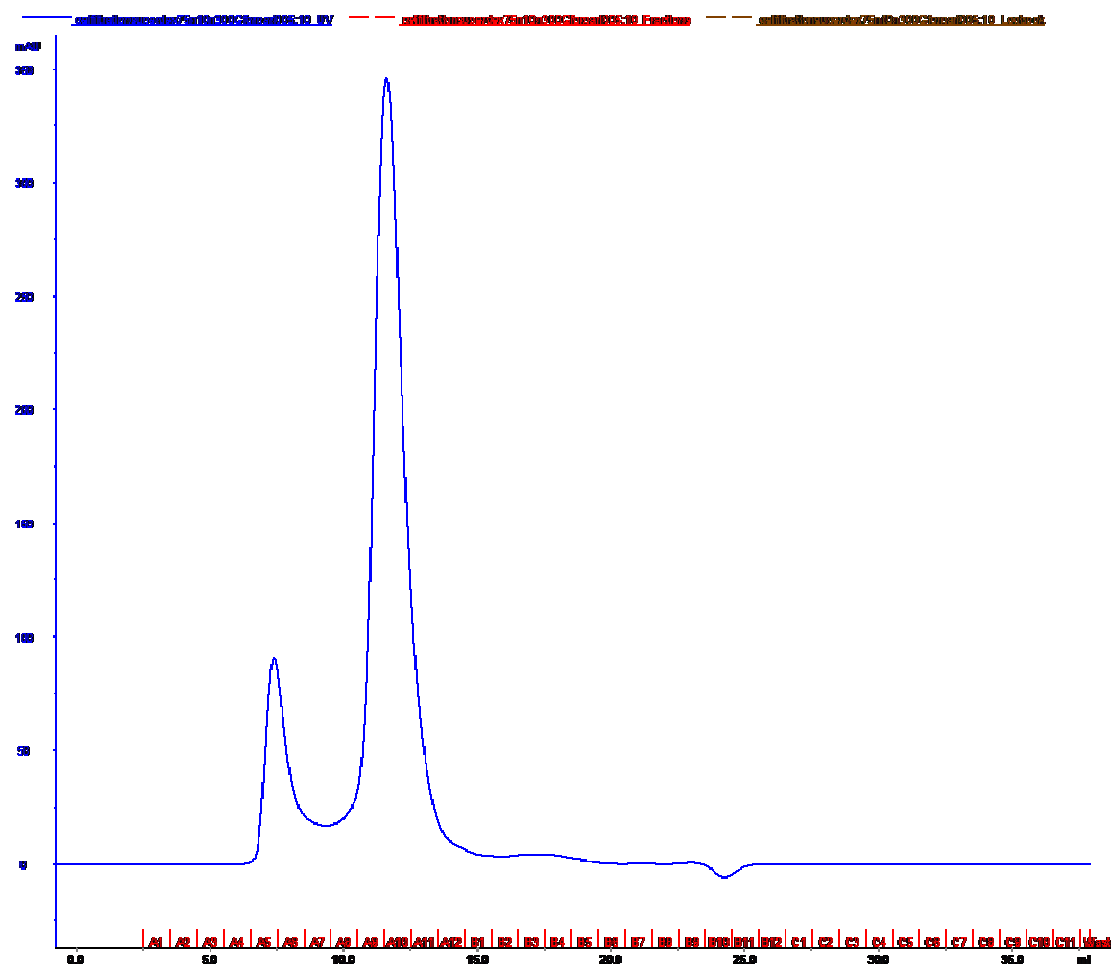
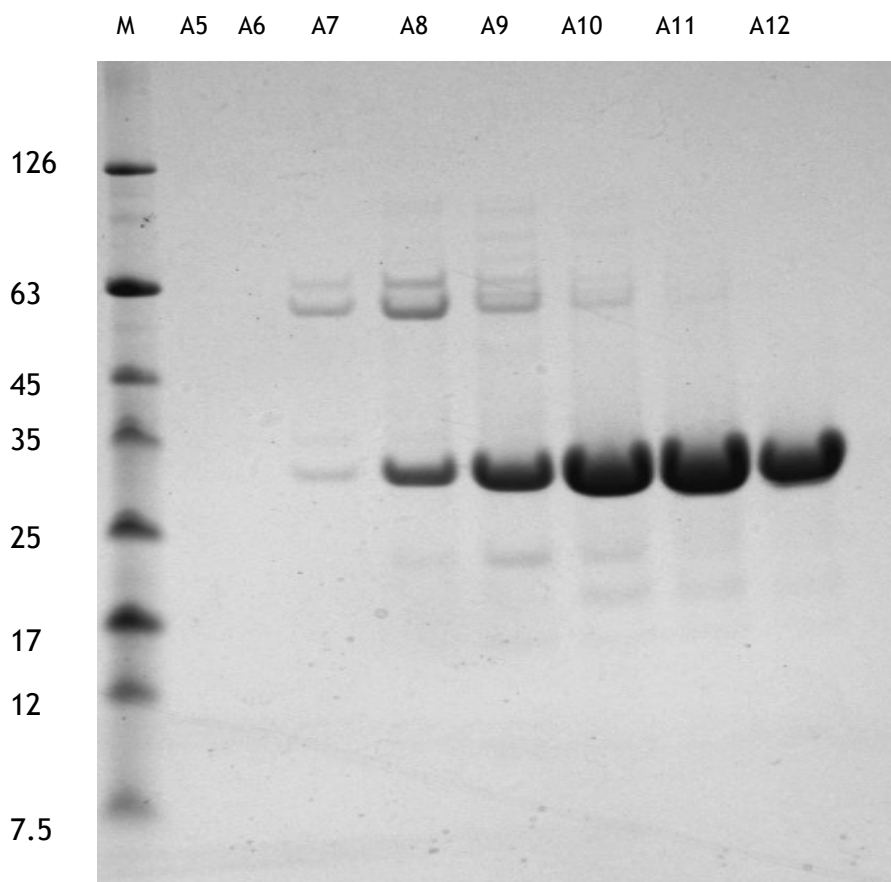


Figure 8.6 Chromatogram for PARI after size exclusion chromatography with individual fractions eluted on molecular size basis



**Figure 8.7 SDS-PAGE analysis for PARI after Ni-purification: M = molecular weight marker in Kda, A5-A7 = elution fraction with impurities, A8-A12 = pure elution fractions containing PARI.**

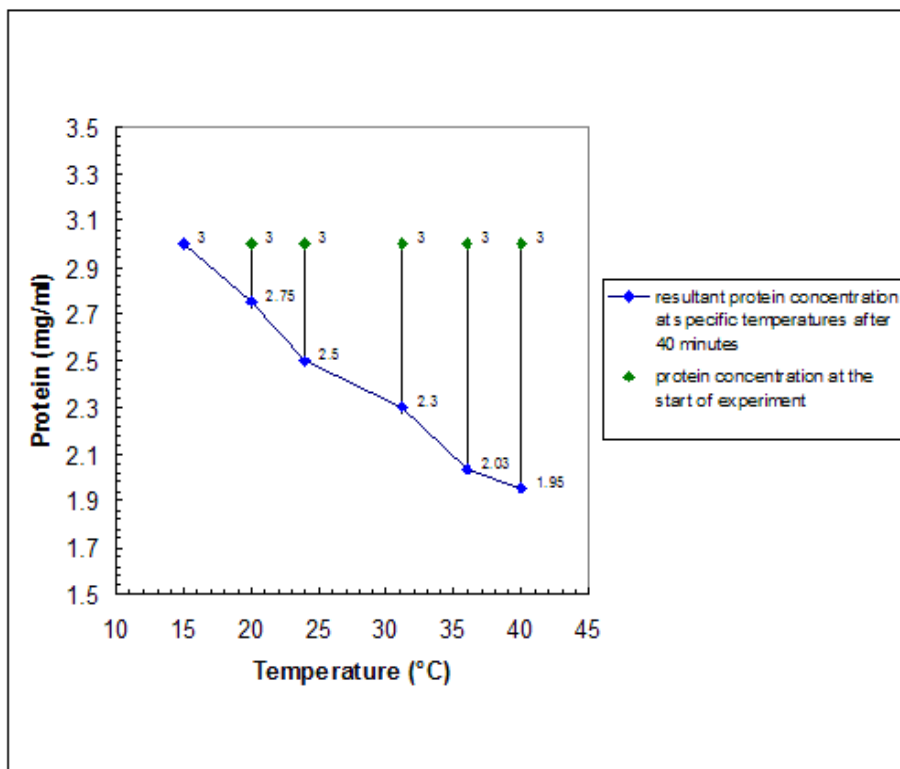
## 8.2 Protein aggregation problems

During the purification, the protein appeared to be prone to precipitation, presumably because of the high Imidazole or NaCl concentration in the elution buffer. Each time when the protein was purified by using Ni-column, the elution fraction went cloudy, and when further subjected to buffer exchange, precipitates started to form; these precipitates were then removed by centrifuging the samples at 3000xg. In order to avoid the precipitation of PARI during its purification different parameters were altered such as

1. Change in pH of purification buffers
2. Varying NaCl concentration in the purification buffers
3. Changing the Tris-HCl concentration in purification buffers

The pH of the buffers was changed as it is possible that if the iso-electric point of the protein coincides with the pH of the solution then the protein will have no overall charge and will precipitate out of solution. Equally the concentration of NaCl, which screens protein charges in solution is known to be able to either increase or decrease solubility which is the basis of salting in or salting out in protein crystallization. Though a slight improvement was observed, but none of the parameters changed were able to completely fix the protein aggregation problem.

A test experiment was carried out to observe the effect of increase in temperature on the aggregation rate of the purified protein. The experiment was carried out by using a 96 well Biometra PCR<sup>®</sup> machine. The starting temperature was set to 15°C and final temperature to 40°C. The starting concentration of the protein was 3mg/ml in 20mM Na phosphate, 50mM NaCl pH: 7.8 buffer. Each of the 100µL protein sample were put in small eppendorfs and then in respective wells with temperatures of 15°C, 20°C, 24°C, 31.2°C, 35.9°C and 40.0°C. The running time for the experiment was set to 40 minutes and the top lid of the instrument was closed to maintain the temperature. It was observed that after 40 minutes the samples at higher temperature went cloudy and when centrifuged at 12000xg for 10 minutes, precipitates were visible in the bottom of the eppendorfs. Further the supernatant was taken and the concentration of the protein sample was measured. The concentration (mg/mL) of individual protein samples after 40 minutes of experiment was measured. After 40 minutes at the following temperatures in the wells 15°C, 20°C, 24°C, 31.2°C, 35.9°C and 40.0°C, the concentration of protein was 3, 2.75, 2.5, 2.3, 2.03 and 1.95 mg/mL respectively. A gradual decrease in the concentration of protein samples was observed with the increase in temperature (Figure 8.8). The experiment indicated that the protein tends to form precipitates as the temperature is increased.



**Figure 8.8** Effect of increase in temperature on the rate of aggregation of PARI, the protein concentration difference before and after the experiment is more visible at higher temperatures (as represented through vertical connected lines)

To see the effect of temperature for continuously longer period of times on the rate of protein aggregation, the experiment was carried out at 20°C and 40°C and the change in protein concentration was measured after each hour for a maximum of 3 hours. The precipitates formed after each hour were removed by centrifuging at 12000xg for 5 minutes, rest of the conditions for the experiment were kept the same as set in previous experiment. The experimental results indicated that the rate of protein aggregation was higher at 40°C than at 20°C (Figure 8.9).

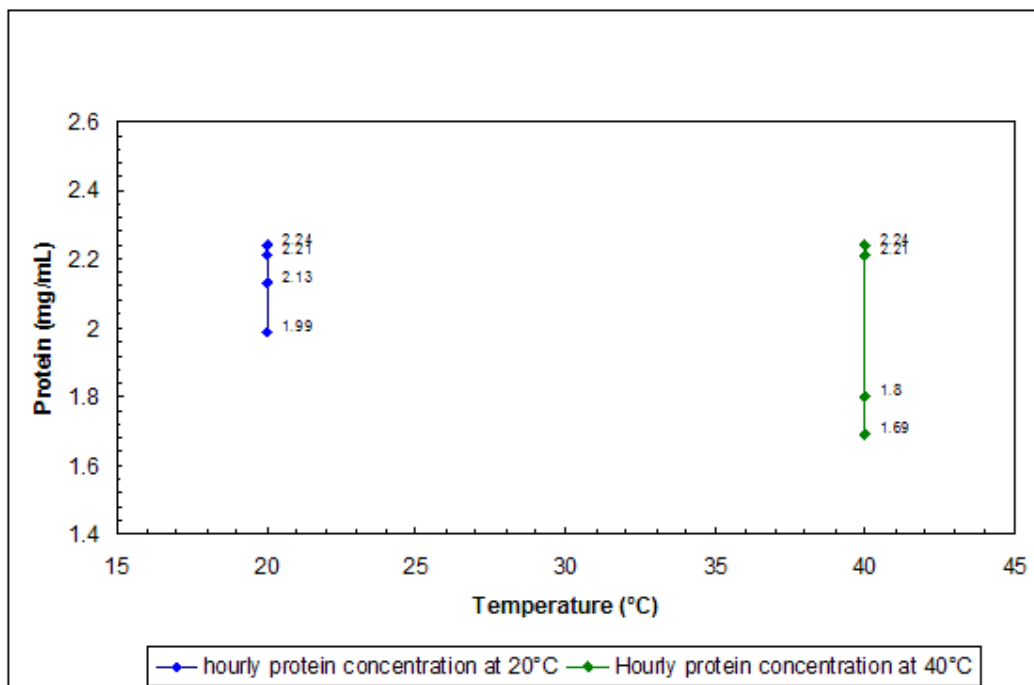


Figure 8.9 Aggregation rate of PARI at 20°C and 40°C for a period of 3 hours (protein concentration was measured after each hour following centrifugation of individual samples, protein concentration at the start of the experiment was 2.24 mg/mL)

From both the experiments it was concluded that

1. The protein is heat labile
2. Unstable at higher temperatures
3. The rate of irreversible aggregation is fastest at higher temperatures

### 8.3 NFGase activity of PARI

PARI NFGase activity was tested by using the established enzyme assay which had been developed for the PAA protein, where hydrolysis of the amide bond is followed at  $A_{210\text{nm}}$  as described in section 2.19.2. The results of the enzyme assays were very surprising as no activity was detected at all. The assays were repeated with varying concentrations of protein and the integrity of the protein was validated by CD and NMR. All possibilities were considered such as misfolding of the protein, proteolysis, inhibition, absence of zinc etc. If PARI was indeed an NFGase or related enzyme, all the possible events would be expected to reduce but not completely eliminate the activity. The sequence of PARI was searched against all protein sequences for the *Ralstonia eutropha* genome sequence at the CMR (Comprehensive Microbial Resource) website. The search



revealed that PARI (2Q7S) is one of a number of genes (HutG1, HutG2 & HutG3) with an NFGase like sequence, including one sequence (HutG1) which is located in the histidine utilization pathway operon and has high sequence similarity to PAA (Figure 8.10). The enzyme assay and analysis of the genome sequences lead to the conclusion that PARI is not a true NFGase and is a different enzyme with different substrate specificity. This was then used as a test case for pharmacophore searching.

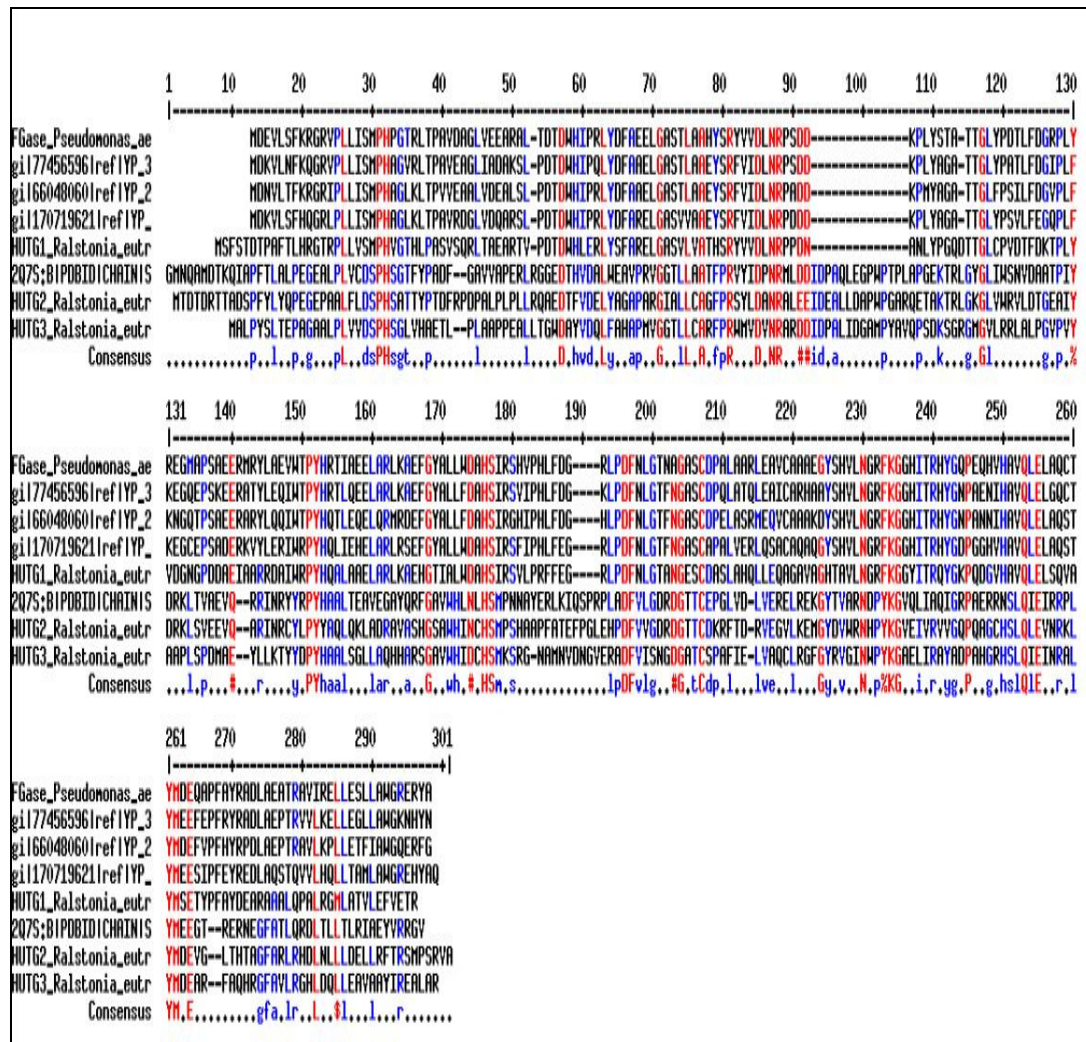
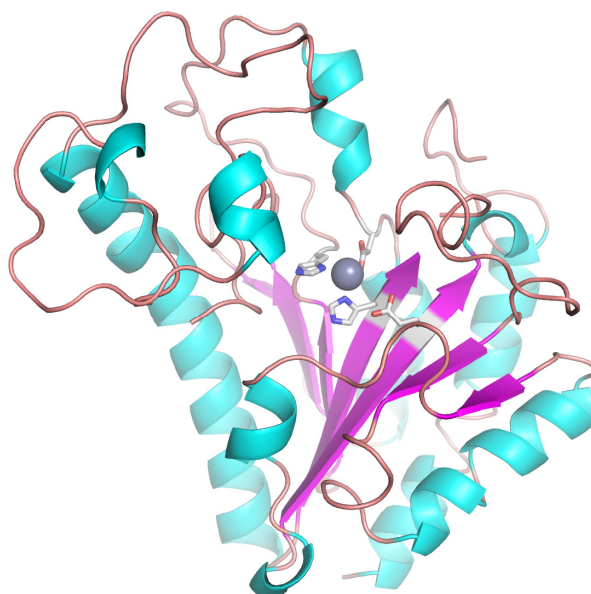


Figure 8.10 Sequence similarity of amino acids between PARI (2Q7S), PAA and some other structural analogues, PAA is the top sequence followed by 4 other NFGase related sequences from *R. eutropha* including PARI (2Q7S), HUTG1 HUTG2 and HUTG3, HUTG1 shows high sequence similarity to PAA while sequences of 2Q7S, HUTG2 and HUTG3 show higher similarity within their structures and form a related but different sub group of proteins, sequences were aligned by using MultAlin {158}.

### 8.3.1 Spasm search

As the PARI and NFGase like proteins have not been characterized with respect to their ligand binding or enzyme mechanism, the enzyme fold does not help in understanding the function by analogy with related proteins. The active site of PARI is reasonably open and contains a zinc binding motif (Figure 8.11).



**Figure 8.11** The structural fold of protein PARI from *R.Eutropha*, the active site  $\text{Zn}^{2+}$  is represented as a grey sphere surrounded by the coordinating residues

In order to search for protein structures with similar arrangement of amino acid residues in the active site, the SPASM {159-161} software was used. This could provide an insight into function as the active site may be conserved while the fold is not e.g. serine protease catalytic triad is found in a number of unrelated protein families.

Initially five amino acid residues, three zinc coordinating residues (HIS31, ASP54 and HIS171) and two conserved residues (ARG78 and GLU249) were selected from the active site of PARI, copied and pasted in to a new window of DSV and then file saved with a .pdb extension. This file was then search against a recent database based on proteins from the PDB by using SPASM software on a local computer as the Web server was no longer available. Certain constraints in

search parameters were optimized by using the available options in the Spasm software, namely:

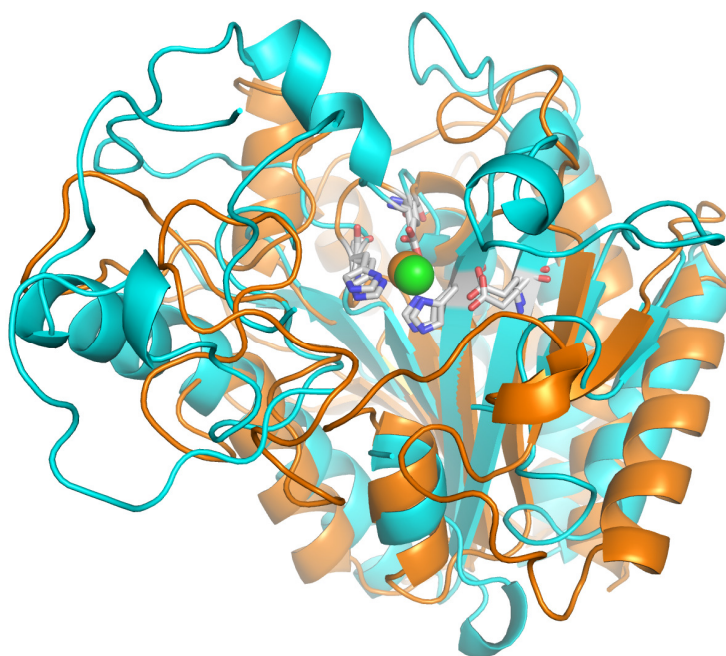
1. Residue substitution options possible
2. sequence directionality not considered and
3. The side chain group would be considered and not the main chain positions.

The above parameters were necessary to optimize as we already knew that there was no protein with a similar fold so the order of the residues and their main chain positions were expected to be different. As a result of SPASM search the numbers of hits obtained were 25 (Table 8.1). Majority of hits were carboxypeptidases while deaminases, oxido reductases and amidases were also included among the hits.

S.No	PDB Code	Type of protein	Organism	Atoms matched	RMSD (Å)
1	1UWY	CARBOXYPEPTIDASE M	Homo sapiens	37	0.61
2	1PCA	PROCARBOXYPEPTIDASE A	Sus scrofa	37	0.63
3	2CTC	CARBOXYPEPTIDASE A	Bos taurus	37	0.63
4	2BO9	CARBOXYPEPTIDASE A4	Homo sapiens	37	0.66
5	1H8L	CARBOXYPEPTIDASE GP180	Lophonetta specularioides	37	0.67
6	1ZLI	Carboxypeptidase B	Homo sapiens	37	0.67
7	1M4L	CARBOXYPEPTIDASE A	Bos taurus	37	0.69
8	2NSM	Carboxypeptidase N catalytic chain	Homo sapiens	37	0.69
9	1CBX	CARBOXYPEPTIDASE A	Bos taurus	37	0.7
10	1JQG	CARBOXYPEPTIDASE A	Helicoverpa armigera	37	0.7
11	5CPA	CARBOXYPEPTIDASE A	Bos taurus	37	0.7
12	1DTD	CARBOXYPEPTIDASE A2	Homo sapiens	37	0.71
13	1Z5R	procarboxypeptidase B	Sus scrofa	37	0.71
14	1NSA	PROCARBOXYPEPTIDASE B	Sus scrofa	37	0.72
15	1OBR	CARBOXYPEPTIDASE T	Thermoactinomyces vulgaris	37	0.73
16	1YW4	Succinylglutamate desuccinylase	Chromobacterium violaceum	37	0.78
17	2O4H	Aspartoacylase	Homo sapiens	37	0.8
18	1KWM	Procarboxypeptidase B	Homo sapiens	37	0.85
19	2BCO	Succinylglutamate desuccinylase	Vibrio parahaemolyticus	37	0.85
20	3B2Y	Metallopeptidase containing co-catalytic metalloactive site	Shewanella denitrificans os217	37	0.86
21	2QVQ	Cellular tumor antigen p53	Homo sapiens	37	0.87
22	2Q4Z	Aspartoacylase	Rattus norvegicus	37	0.94
23	2G9D	Succinylglutamate desuccinylase	Vibrio cholerae	37	1.05
24	1JWQ	N-ACETYLMURAMOYL-L-ALANINE AMIDASE	Paenibacillus polymyxa	37	1.37
25	1XOV	Ply protein	Bacteriophage psa	37	1.49

**Table 8.1 Hits obtained from SPASM search along with their PDB codes, type of protein and source of origin, most of the hits were carboxypeptidases**

The hits could be transformed onto the coordinates of PARI using a transformation matrix which was applied using Coot {95}. However, due to differences in convention the Spasm rotation matrix is the inverse of that used by coot so the matrix had to be transposed. Good agreement between the residues involved in the zinc binding were seen between carboxypeptidase and PARI, also other residues were found in common which could be included in the alignment of the proteins. Among other carboxypeptidases structures that had an active site architecture similar to PARI, 1CBX was found to be the best hit (Figure 8.12). The 1CBX structure was determined with a ligand in the active site and this could provide very useful starting point for pharmacophore searching.



**Figure 8.12 Superimposed folded structure of PARI (blue) with 1CBX CarboxypeptidaseA (orange). The superimposed structures demonstrate the overall fold similarity and identical orientation of the active site residues (the active site residues around the zinc atom are shown as stick model)**

## 8.4 Pharmacophore searching for PARI

All the pharmacophores were generated by using the query atom method as described in detail in section 3.5.3.

### 8.4.1 Generation of pharmacophore for PARI

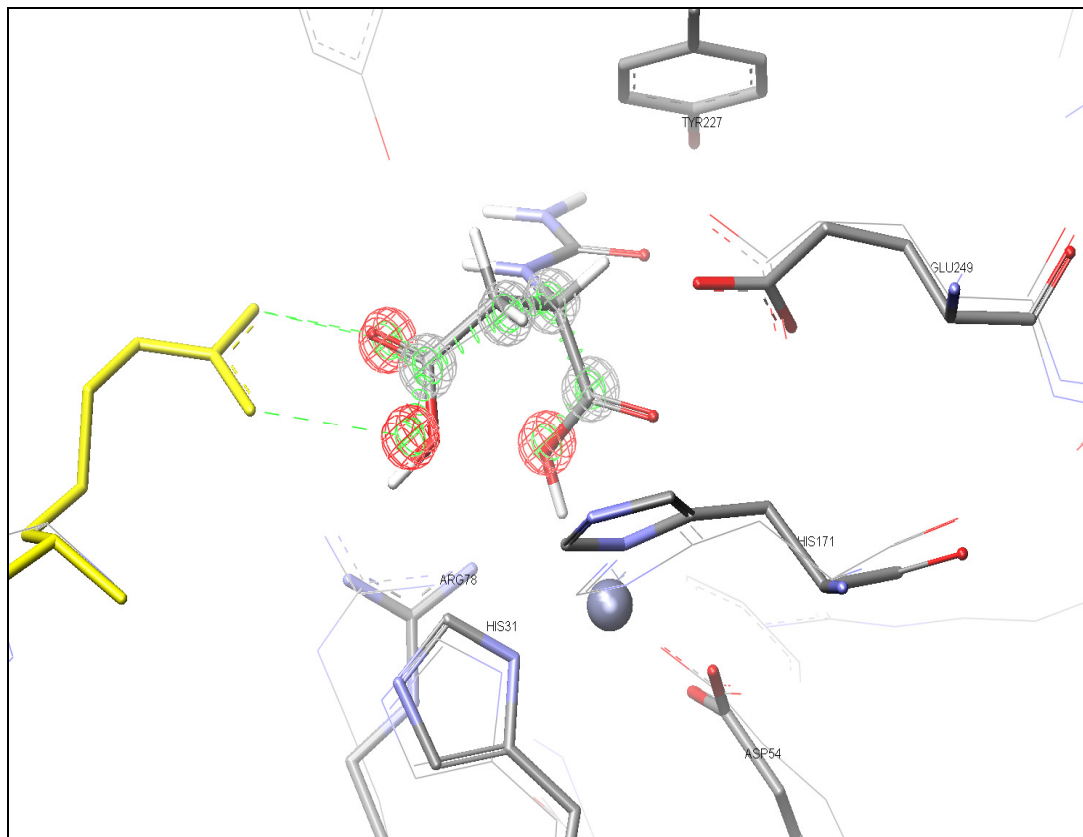
All the hits obtained through SPASM search were visualized in COOT®. 1CBX was selected as a template model on the bases of close resemblance of active site residues to PARI. The structures of 1CBX and PARI were visualized simultaneously in COOT® by using the superimposing feature. The x, y and z co-ordinates of the hits were transformed with reference to PARI by using the operator function (through the route of calculate, scripting) and python command:

```
transform_molecule_by (molecule number) m11 m12 m13 m21 m22 m23 m31  
m32 m33 x y z
```

The transformed coordinates were then saved with a .pdb extension and further visualized in DSV®. Both the PARI and 1CBX structures were optimised by using the structure superimpose and add selecting pairs of atoms to act as tethers option. This way an optimal alignment was made possible.

#### 8.4.1.1 Pharma PARI 1

Clearly, the inhibitor for 1CBX would not be a suitable inhibitor of the enzyme as the benzyl group is not present exactly in active site pocket near the zinc atom, however the key aspects of the ligand recognition, i.e; the carboxylate group interactions with the zinc ion could be utilized. Alignment of 1CBX with PARI enabled to take the key conserved features of the inhibitor L-benzyl succinate found in the active site of 1CBX and transfer them to the PARI active site. The pharmacophore with a simple minimal description along with exclusion spheres was generated. The generated pharmacophore was searched through the ChEBI.bdb database. The search gave 76 hits in just 40 minutes time. The hits included notably benzyl succinic acid, succinic acid, L-aspartic acid and N-carbamoyl-L-aspartic acid (Figure 8.13).



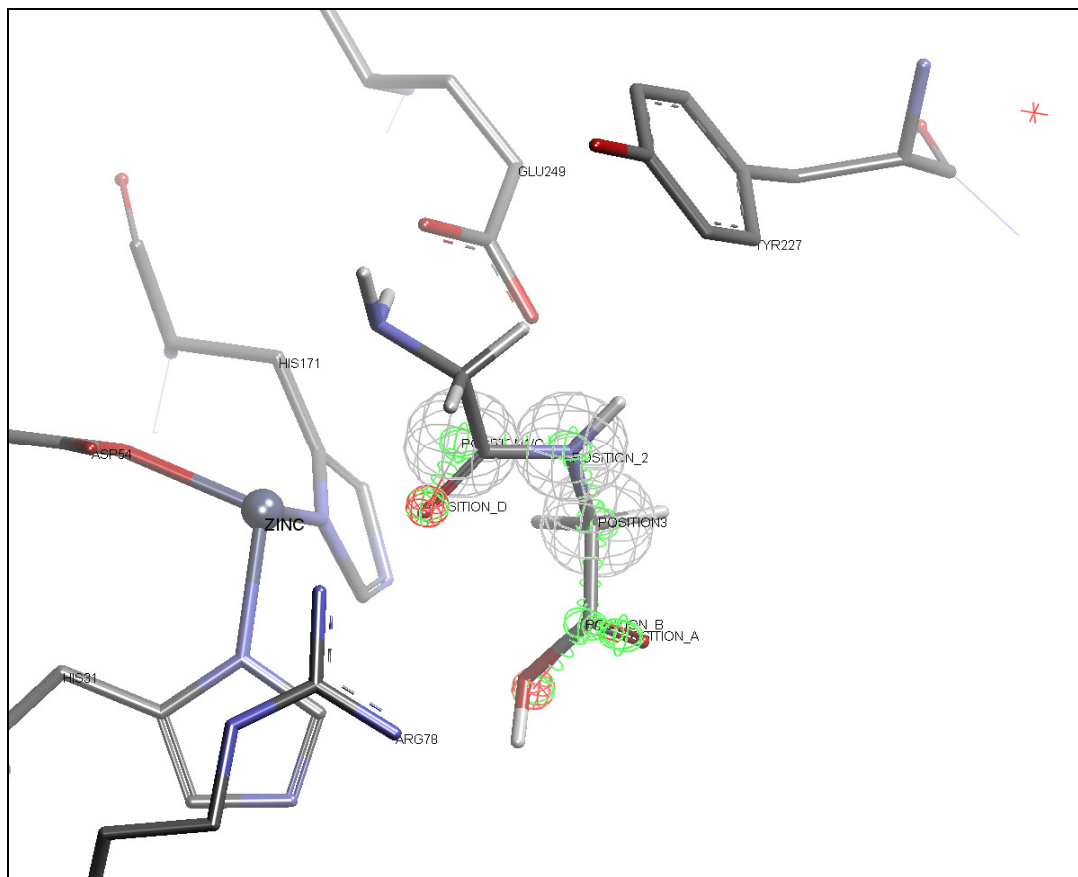
**Figure 8.13** Superimposed structures of ICBX (line models) and PARI (stick model) along with the pharmacophore containing N-carbamoyl L-Aspartic acid as a hit in the active site surrounding by conserved residues, green dashed line shows crucial H-bonds between ARG (highlighted in yellow which is slightly disordered in PARI) and carboxy terminus of the ligand all exclusion spheres, some amino acid residues and water molecules have been removed for clarity viewing

#### 8.4.1.2 Pharma PARI 2

All the hits obtained through first pharmacophore had only carbon atoms specified for position 1, 2 and 3 in the query ligand. In order to see the effect on the chemical features of the hits, multi atoms were specified on these positions including C, N and O. The radius of uncertainty location sphere around positions 1, 2 & 3 was changed from 0.4Å to 0.8Å (Figure 8.14). Further the H-count for position A and C was changed to 0 and bond order between A-B and C-D was set to single/double. The pharmacophore was then searched through the ChEBI.bdb database. The search gave 126 hits, in which some of the hits had nitrogen atom at position 2 (which is important for catalytic activity of a Carboxypeptidase, in terms of cleaving the peptide bond). The type of hits obtained demonstrated the benefit of relaxing/optimization of the above mentioned constraints, thus



getting more potential ligands. The hits also included 5 dipeptides, geometrically positioned in a favorable manner (for peptide cleavage) preferred in the active site of a Carboxypeptidase. The hits in particular included N-carbamoyl-L-aspartic acid, N-carbamoyl-L-valine, benzyl succinic acid, glycylglycine, L-cysteinylglycine and L-prolylglycine



**Figure 8.14** Pharmacophore with the size of the uncertainty location sphere at position 1, 2 & 3 changed from 0.4Å to 0.8Å, among the hits glycylglycine (a dipeptide) is represented as a stick model superimposed on query ligand in the active site of the protein, the presence of nitrogen atom at position2 favours the possible cleavage of peptide bond. (All exclusion spheres, certain amino acid residues and water molecules have been hidden for clarity viewing)

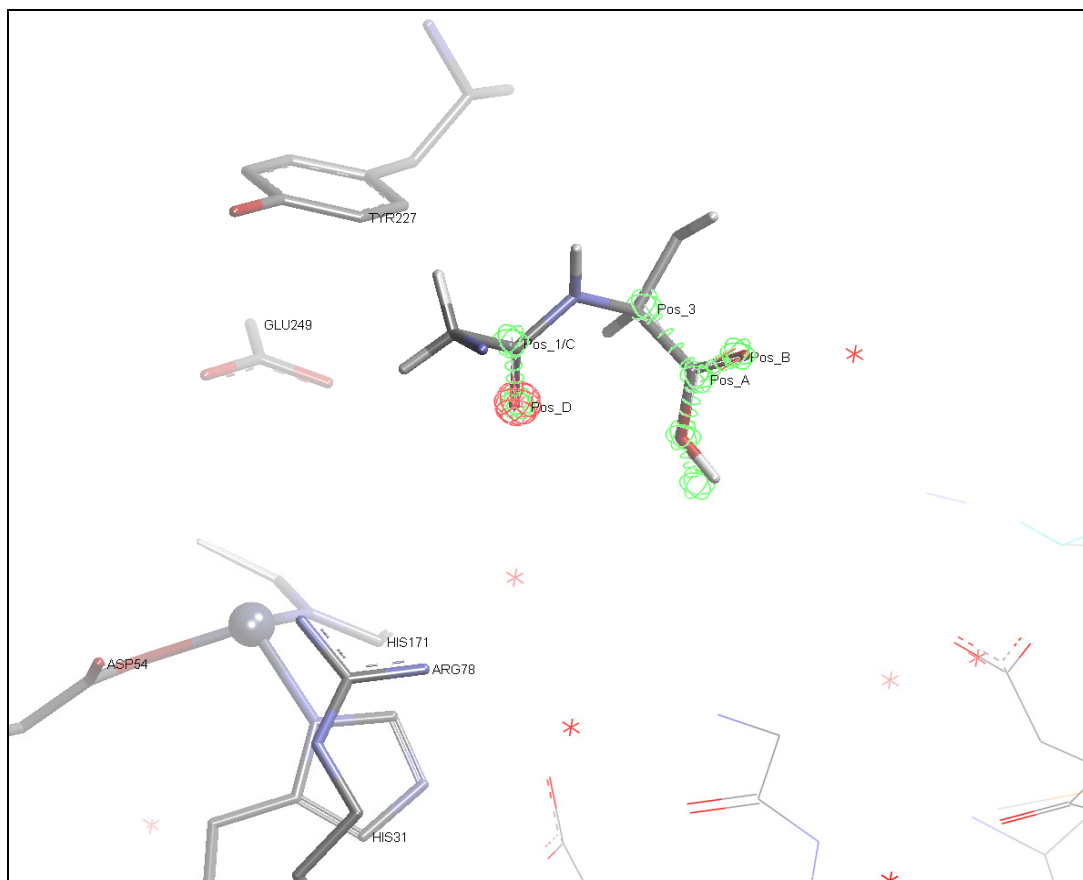
To check the number of peptides in the database a chemical searching of the database with an alanyl group gave only 47 hits which included only Ala-Glu as a dipeptide. It became clear that very few peptides (di/tri) were found in the database. In order to fulfil this gap a peptide database was generated (as described in section 3.4.2). The generated database had all possible combinations of di and tri-peptides i.e. 8,400 compounds and was used in later pharmacophore searches.

### 8.4.1.3 Pharma PARI 3 (sub-structure fragment based search)

In order to avoid a biased search via a ligand model, the position 2 was removed from the previous pharmacophore and the location of the C $\alpha$  carbon was not defined. A minimal pharmacophore was specified which contained:

1. Position C along with the carbonyl oxygen at position D towards the zinc metal ion.
2. Carboxyl carbon at position A and its corresponding two oxygens towards the ARG78 of the protein.

The modification left a fragment type sub structure pharmacophore in place (Figure 8.15). The resultant pharmacophore was then searched against the dipeptide and tripeptide database.

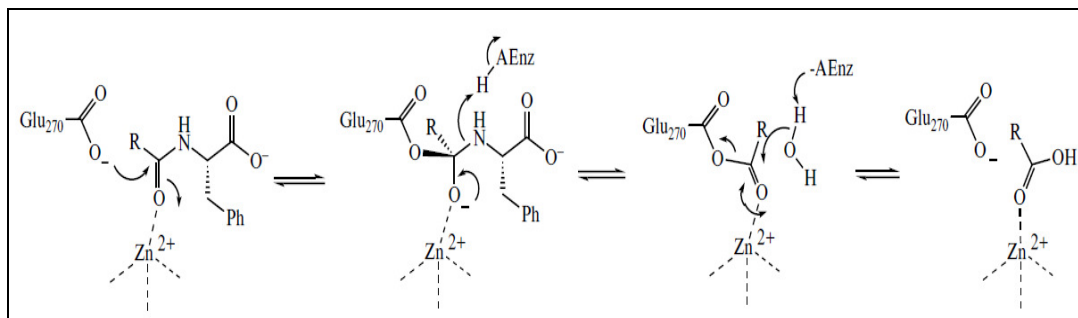


**Figure 8.15** Position\_2 removed from the pharmacophore, leading to a substructure fragment style pharmacophore. Asp-Gly-Ser a tripeptide among the hits is represented as stick model (All exclusion spheres, certain amino acid residues and water molecules have been removed for clarity viewing)

Interestingly 200 hits were obtained through the search, majority of the dipeptides and tripeptides had their C-terminal positioned towards the GLU249



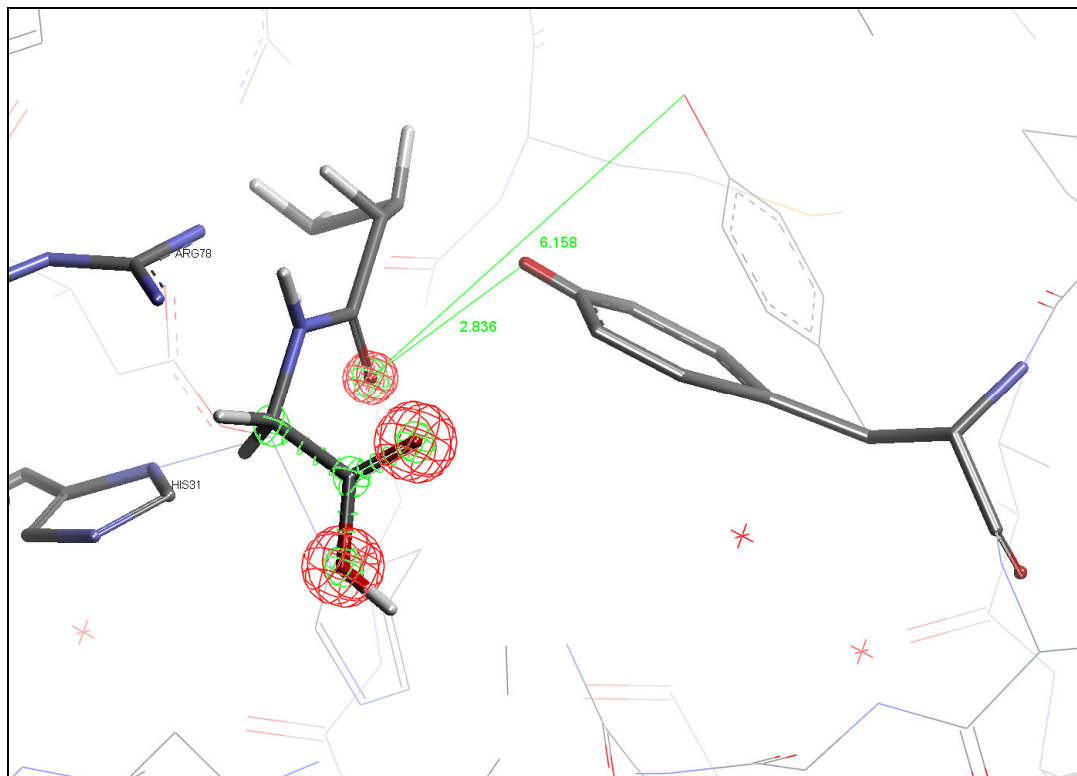
as per proposed mechanism for a carboxypeptidase (Figure 8.16) except some peptides with aspartate which had the R-group oriented towards the carboxyl binding site. The hits included 101 dipeptides like Ala-Gln, Ala-Gly, Ala-Ile, Ala-Pro, Ala-Ser and 99 tripeptides like Asn-Ala-Asn, Asn-Ala-Pro, Asn-Ala-Thr, Asn-Gly-Asp, Asn-Gly-Gly. Due to the absence of well defined hydrophobic pocket, peptides (di/tri) with bulky hydrophobic residues were absent amongst the hits.



**Figure 8.16** Schematic representation of possible nucleophilic mechanism for Carboxypeptidase A, R= peptide chain, adapted from{162}.

#### 8.4.1.4 Pharma PARI 4

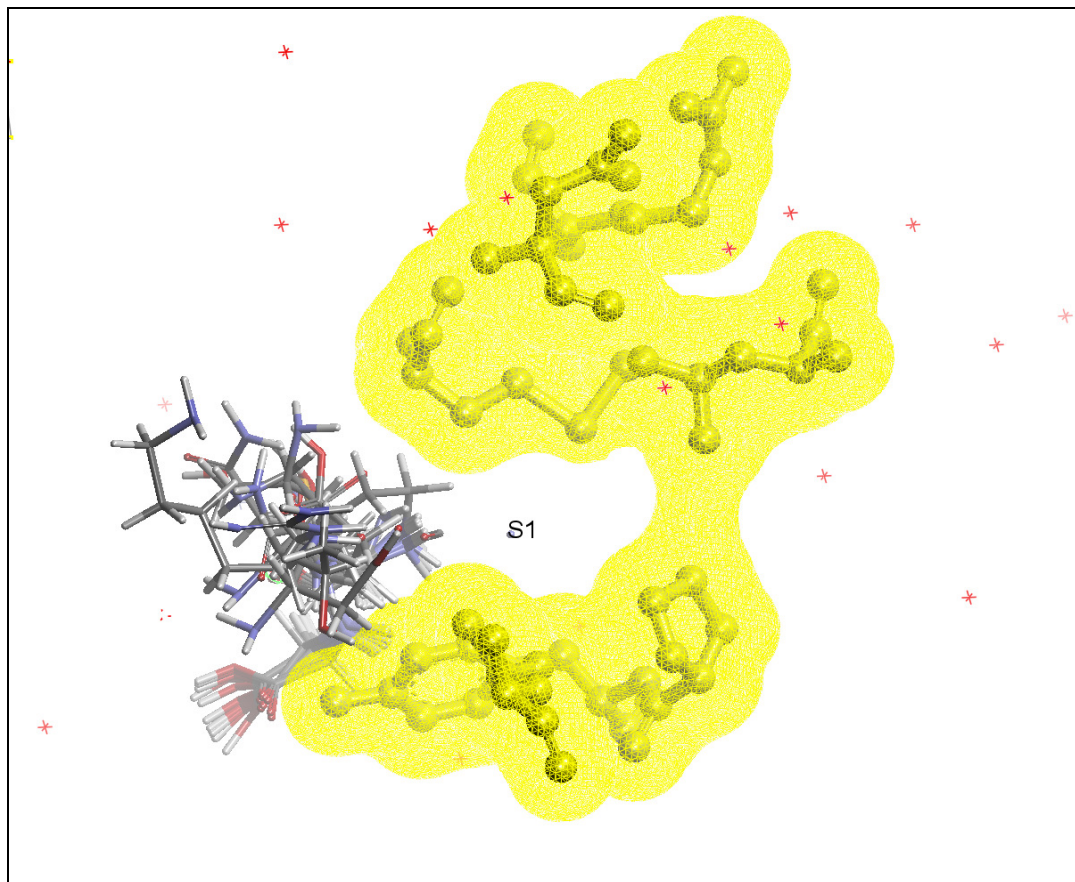
In a proposed mechanism of carboxypeptidases {163}, tyrosine residue has a significant role in the stabilization of the substrate. Upon the introduction of the substrate tyrosine closes in to the active site ( $S1^{\wedge}$ ) forming H-bonds with the carbonyl oxygen of the peptide (Figure 8.17) and thus creating a pocket ( $S1$ ) to fit in the whole substrate (either a di or tripeptide). In order to identify potential residues that could bind in the  $S1$  binding site the previous pharmacophore was modified by changing the position of the TYR227 (by using the Torsion feature of the DSV®). The resultant distance between the carbonyl oxygen of the query ligand and TYR227 (previously 6.158Å) was optimized to a good H-bond distance of 2.836Å (Figure 8.17). The resultant pocket on top of TYR227 was visualized by using the solvent surface feature of DSV®, which was mostly hydrophobic in nature, comprising of MET173, PRO226, ARG251, ASN175, LEU189 (towards the bottom of the pocket) and LEU181 (towards the top of the pocket). The exclusion spheres were introduced around the optimized active site and the resultant pharmacophore was searched through the di and tri peptide database. The search gave 26 hits which included 8 dipeptides like Cys-Gly, Glu-Gly, Gly-Gly, Pro-Gly and 18 tripeptides like Ala-Gln-Gly, Ala-Thr-Gly, Arg-Ser-Gly, Asp-Gly-Gly, Asp-His-Ser. All the hits demonstrated the formation of H-bond between their carbonyl oxygen and TYR227 of the enzyme.



**Figure 8.17** Representation of the change in position of the TYR227 at a H-bond distance from POS\_D, the final position of TYR227 is shown as stick model while the previous position is in line model, Val-Gly a dipeptide among the hits is represented as stick model (all exclusion spheres, certain amino acid residues and water molecules have been removed for clarity viewing)

#### 8.4.1.5 Pharma PARI 5

The orientation of hits towards the hydrophobic pocket (S1) was not as desired, in order to get hits fitting well within the hydrophobic pocket, the size of the uncertainty location sphere around the carboxyl carbon (Pos\_1/C) was changed from 0.8Å to 0.6Å, the rest of the pharmacophore was kept the same. The search gave 11 hits from the dipeptide/tripeptide database, but still the R-groups of the hits were not fitting in the hydrophobic pocket (Figure 8.18).



**Figure 8.18** Graphical representation of the hits (in line model) in the active site of the protein, the hydrophobic pocket (labelled as S1) is empty while the R groups of the hits are located outside the hydrophobic pocket.

## 8.5 Ligand selection and synthesis

Given the difficulty in assigning specific amino acid groups to the S1 site and the potential variation in possible peptides, it was decided to try and identify binding ligand to the active site (S1<sup>^</sup>, C-terminal carboxylate binding site). Simple amino acids as products are likely to have weak affinity. Therefore N-carbamoyl amino acid derivatives would seem to be suitable candidates as they cannot be hydrolysed by the enzyme and should give insight regarding affinity of the protein for different amino acid derivatives.

Carbamoyl derivatives of different amino acids like carbamoyl cysteine, carbamoyl lysine, carbamoyl glutamate, carbamoyl glutamine, carbamoyl phenylalanine, carbamoyl methionine, carbamoyl valine, carbamoyl threonine, carbamoyl leucine, and carbamoyl-histidine were synthesized by using the microwave method as used by Verado *et al* [164]. All the reactions were carried

out in a domestic microwave oven (Panasonic NN-SD466M) at 1000W. Mass spectra were obtained by using a JEOL JMS-700 spectrometer.  $^1\text{H}$  NMR and  $^{13}\text{C}$  spectra were recorded on a Bruker DPX-400 spectrometer with chemical shift values in ppm relative to TMS ( $\delta_{\text{H}}$  0.00 and  $\delta_{\text{C}}$  0.0) as standard with  $\text{D}_2\text{O}$  as a solvent at room temperature. Specific optical rotations of the compounds were determined in 0.25M NaOH solution in methanol at  $20^\circ\text{C}$  (compound concentration 15mg/mL) by using Automatic polarimeter APV-6W, RUDOLPH RESEARCH ANALYTICAL, NJ, USA. All the experimental data (optical rotation, mass spectrometry,  $^1\text{H}$  NMR and  $^{13}\text{C}$ ) was consistent with the already published data {164}. The details of the microwave procedure for carbamoyl glutamate synthesis are given below.

### ***8.5.1 Synthesis of Carbamoyl Glutamate:***

Carbamoyl glutamate was synthesized by the Microwave-assisted method {164}. L-glutamic acid (6.0 mmol) was added in to a 50 mL beaker containing a stirred solution of NaOH (0.24g, 6.0mmol) in water (1.8mL). When a clear solution was obtained, urea (0.62g, 10.30mmol) was added, and the reaction mixture was made homogenous by well mixing with the help of a glass rod. The reaction vessel was covered with a watch glass and placed into an unmodified domestic microwave oven, together with a 500mL beaker containing water (300–400mL), and irradiated for 4 minutes at 1000W. The reaction mixture was then cooled at  $0^\circ\text{C}$ , and 6M HCl (1.0mL, 6.0mmol) was added while stirring. The obtained solid was filtered and washed with water (6.0mL) to eliminate excess urea and NaCl, yielding carbamoyl glutamate in a spectroscopically pure form. The % yield of the synthesized compounds was 75.2%.

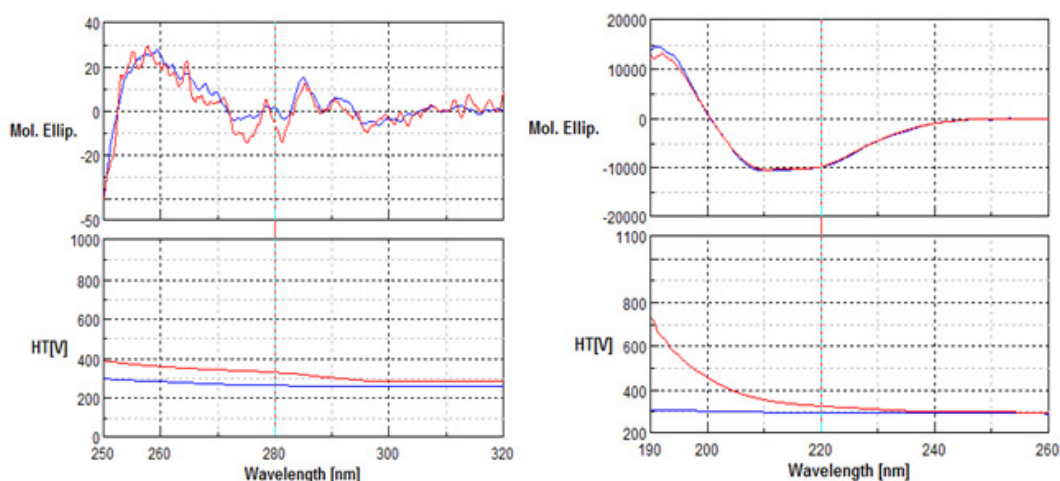
Spectroscopic data was in accordance with the literature.  $\{\alpha\}_{\text{D}}$  +4.2; (m/z) (CI) 191.25 (95%), 164.23 (35%), 120.22 (15%), 97.22 (20%), 71.15 (45%);  $\delta_{\text{H}}$  (400 MHz,  $\text{D}_2\text{O}$ ) 1.81–1.92 (m, 2H, 3- $\text{CH}_2$ ) 2.03–2.13 (2H, m, 2- $\text{CH}_2$ ), 4.09 (q,  $j$  5.0, 4-CH);  $\delta_{\text{C}}$  (100MHz,  $\text{D}_2\text{O}$ ) 29.15, 34.16, 55.70, 161.0, 180.20 and 182.55.

Brenda enzyme database (<http://www.brenda-enzymes.info/>) was used as a reference to add a sensible concentration of potential ligands to the protein. The affinity values for different ligands in case of carboxypeptidases were in the range of 1–2mM.

## 8.6 CD results for PARI

To see the effect of potential ligands on the secondary structure of the protein, particularly on the aromatic regions of the protein the selected potential ligands were added to the protein and their near and far UV spectra were recorded. All the CD spectra were recorded at a protein concentration of 1.7mg/mL; cell path length was 0.01cm and 0.2cm for Far and Near UV analysis respectively. The CD experiments were carried out in buffer solution of 20mM Na-phosphate, 50mM NaCl pH: 7.8 at 20°C.

The near CD spectrum for PARI with carbamoyl cysteine (Figure 8.19) demonstrated subtle changes in the 270-290nm range, which may suggest some slight structural rearrangements in the binding site.



**Figure 8.19** Near UV (250-320nm) and Far UV (190-260nm) CD spectra for PARI (blue) and PARI with 20mM Carbamoyl cysteine (red).

The spectral changes brought by the addition of carbamoyl lysine, carbamoyl glutamate and L-Histidine (Figure 8.20-22) were very subtle and the pattern observed for these ligands was more or less the same in the Near UV region. This similarity in CD spectra might indicate that these ligands are causing the same effect in or around the binding site of the protein. Pertaining to the fact that the ligands are small and thus are unlikely to cause a big shift in the near UV or far UV spectra, the overall composition of  $\alpha$ -helices and  $\beta$ -strands nearly remained the same in case of these ligands.

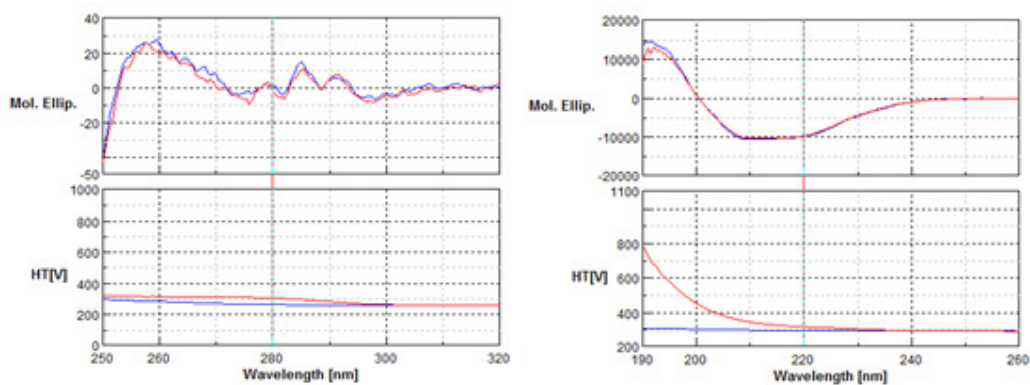


Figure 8.20 Near UV (250-320nm) and Far UV (190-260nm) CD spectra for PARI (blue) and PARI with 20mM Carbamoyl lysine (red).

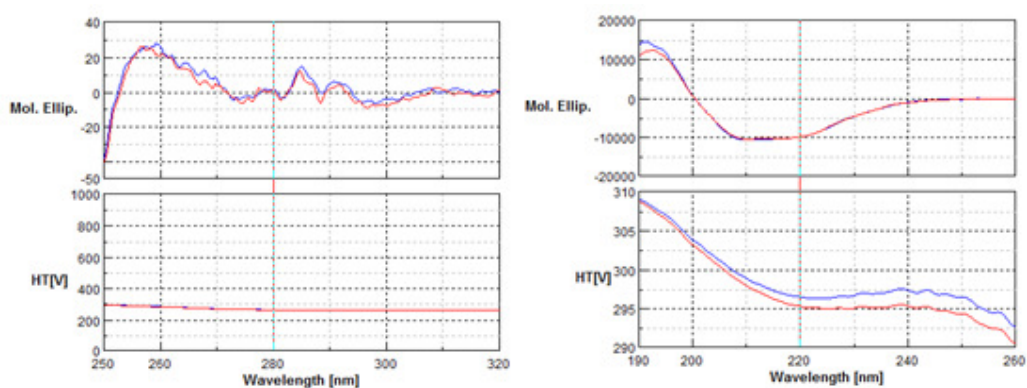


Figure 8.21 Near UV (250-320nm) and Far UV (190-260nm) CD spectra for PARI (blue) and PARI with 20mM Carbamoyl glutamate (red).

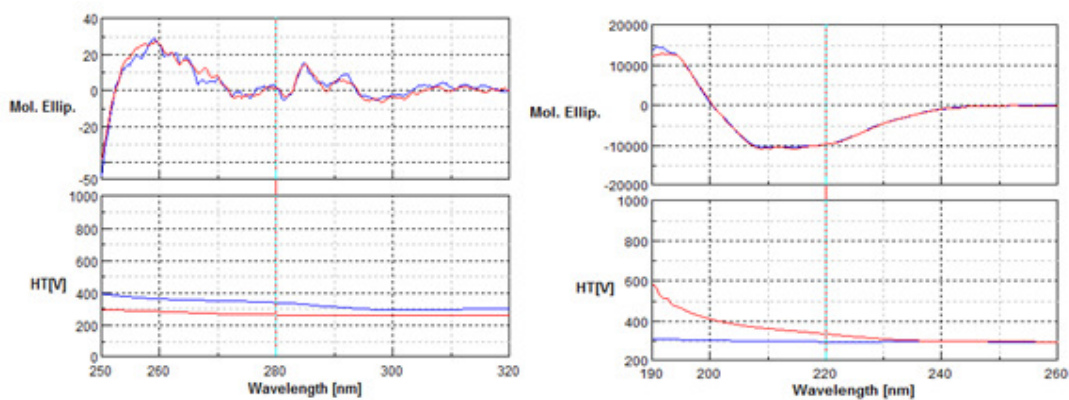
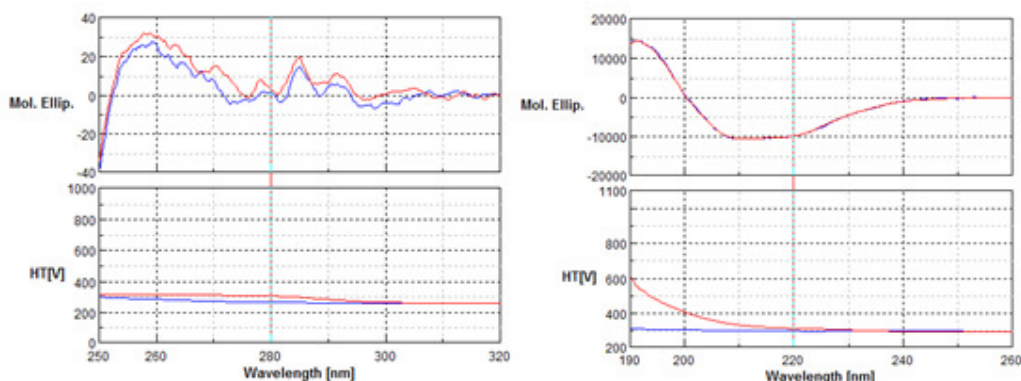


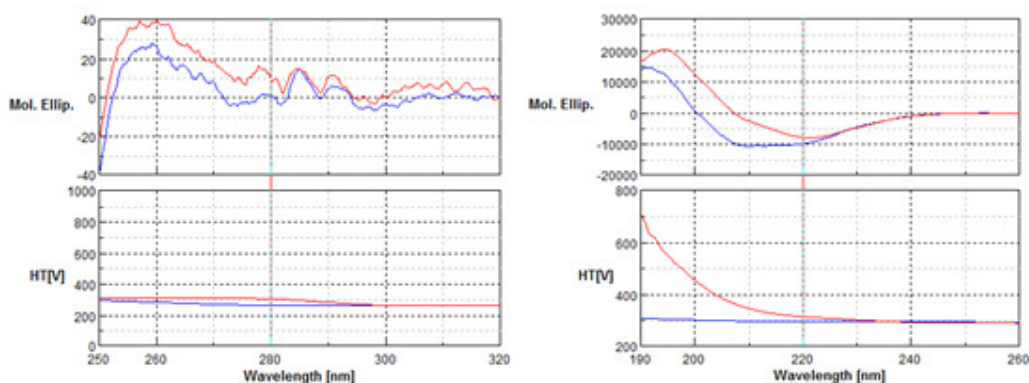
Figure 8.22 Near UV (250-320nm) and Far UV (190-260nm) CD spectra for PARI (blue) and PARI with 20mM L-Histidine (red).



The spectral changes observed for L-glutamate and N-formyl L-glutamate were more significant than the rest of the ligands (Figure 8.23-24). The changing pattern in the near UV region (particularly in 255-310nm range) for L-glutamate and N-formyl L-glutamate was pretty similar, though the shifts were much bigger in case of N-formyl L-glutamate. In the far UV region changes were observed for N-formyl L-glutamate (particularly in 190-220nm range). The bigger shifts both in the near UV and far UV region due to the addition of N-formyl L-glutamate suggests some changes in the secondary structure of the protein.



**Figure 8.23** Near UV (250-320nm) and Far UV (190-260nm) CD spectra for PARI (blue) and PARI with 20mM L-glutamate (red).



**Figure 8.24** Near UV (250-320nm) and Far UV (190-260nm) CD spectra for PARI (blue) and PARI with 20mM N-formyl L-glutamate (red).

By comparing all the CD spectra both in the near UV and far UV region, it appears that N-formyl L-glutamate brings more significant changes than any other ligand. The effect of different ligands on the protein in descending order can be given as

**N-formyl L-glutamate > L-glutamate > carbamoyl cysteine > carbamoyl  
histidine = carbamoyl lysine = carbamoyl glutamate**

## **8.7 NMR experiments on PARI**

In order to see the effect in terms of ligand binding, HSQC spectra were carried out by using the  $^{15}\text{N}$ -labelled protein. The protein was over expressed in M9 media supplemented with  $^{15}\text{N}$ -labelled  $\text{NH}_4\text{Cl}$ . The protein samples for HSQC spectra both with and without the ligand were prepared in the same way as described in section 6.6.1. The ligands were prepared in the same buffer solution as of the protein and further the pH of the ligand solution was adjusted to the protein solution. To get a cleaner spectra, and better signal to noise ratio, the experiments were carried out at different temperatures. The spectra were recorded with  $5^\circ\text{C}$  increase in temperature, starting from  $20^\circ\text{C}$  up to  $40^\circ\text{C}$ .

The increase in temperature with the aim to get higher signal to noise ratio resulted in sharp signals for the protein. Initially the HSQC spectra for the protein were recorded in buffer solution of 20mM Na-phosphate, 150mM NaCl pH: 7.5. The appearance of the HSQC spectra was not very clear, therefore to increase the signal to noise ratio the salt concentration was decreased to 50mM in the buffer solution which resulted in sharp peaks for individual residues. All the experiments were carried out at  $35^\circ\text{C}$  in buffer solution of 20mM Na-phosphate, 50mM NaCl pH: 7.8. The final concentration of protein and different ligands was  $100\mu\text{M}$  and 20mM respectively.



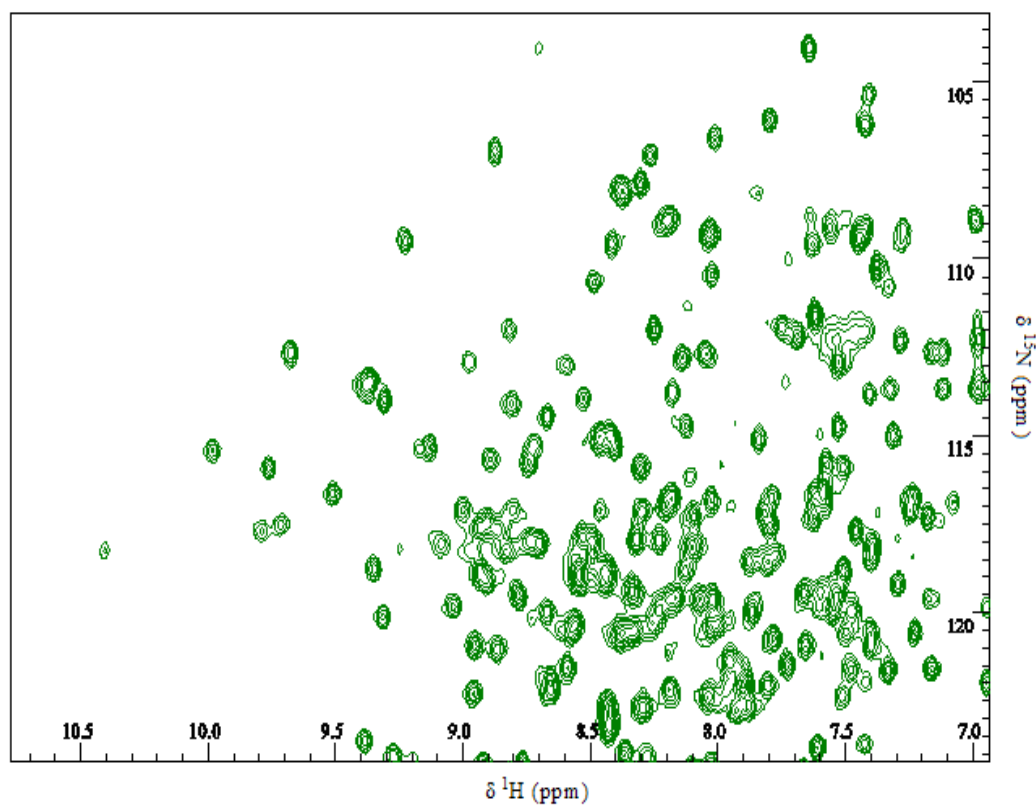
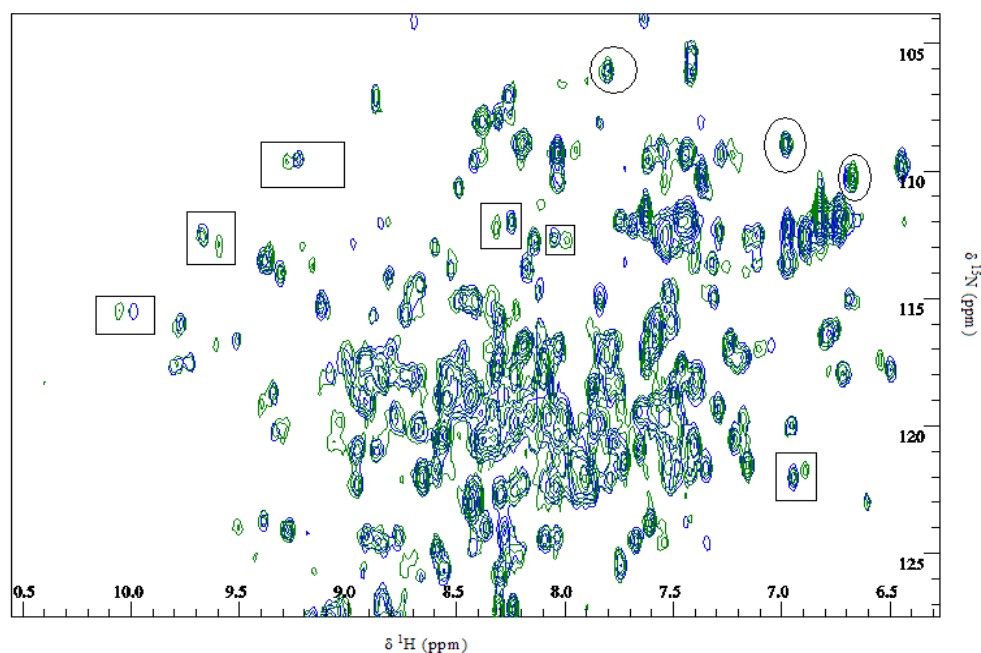


Figure 8.25 NMR-HSQC spectra for  $^{15}\text{N}$ -labelled PARI without any ligand

### 8.7.1 N-formyl L-glutamate with PARI

The overlaid HSQC spectra for the protein along with N-formyl L-glutamate showed some marked changes. It was observed that certain peaks significantly changed their positions while few peaks disappeared and some new peaks appeared as well (Figure 8.26). The changes brought about by the addition of N-formyl L-glutamate were more profound and visible than any other ligand. The changes can be attributed to a direct interaction between certain amino acid residues and N-formyl L-glutamate.



**Figure 8.26** Overlaid HSQC spectra for PARI along with 20mM N-formyl L-glutamate, empty square boxes around peaks represent the change in position of peaks due to the addition of ligand, empty circles around peaks represent no change due to the addition of ligand (green spectra = protein alone, blue spectra = protein with ligand)

### 8.7.2 Carbamoyl histidine with PARI

More marked changes in the peaks positions were observed due to the addition of carbamoyl histidine to the protein, it appeared that some peaks travelled a longer distance and appeared at a new position while some peaks slightly changed their position. The changing pattern occurring due to the addition of carbamoyl histidine had some similarities with N-formyl L-glutamate. The bigger chemical shift for certain residues was the same as due to addition of N-formyl L-glutamate (Figure 8.27).

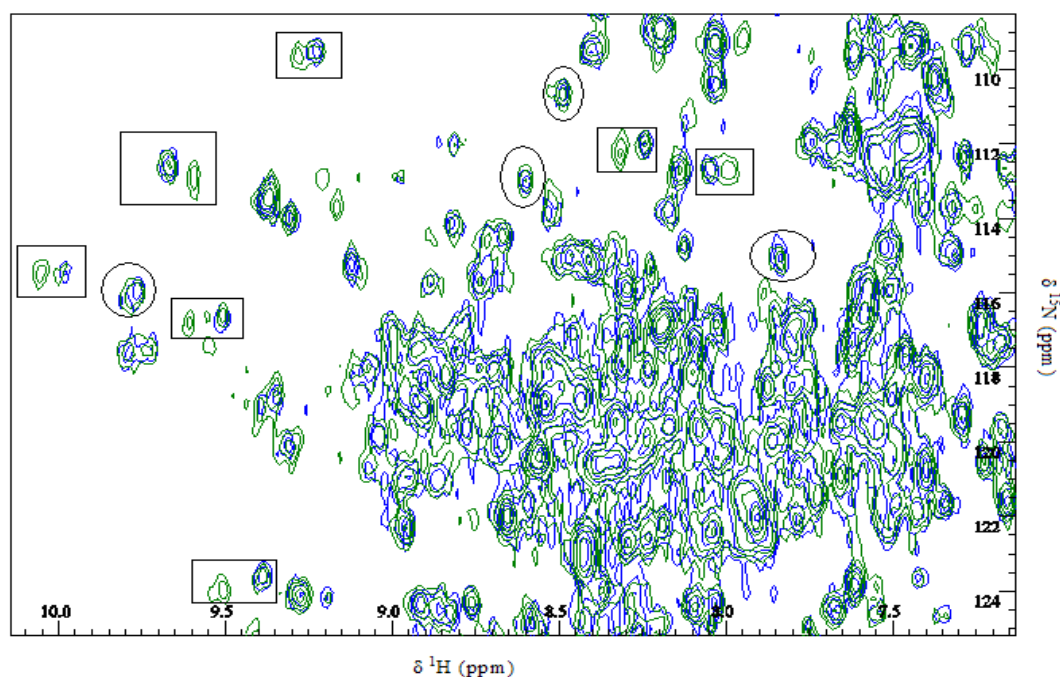


Figure 8.27 Overlaid HSQC spectra for PARI along with 20mM carbamoyl histidine, empty square boxes around peaks represent change in position of peaks due to the addition of the ligand, empty circles around peaks represent no change due to the addition of ligand (green spectra = protein alone, blue spectra = protein with ligand)

### 8.7.3 L-glutamate with PARI

The overlaid HSQC spectra for protein along with L-glutamate indicated significant changes. Some peaks completely disappeared due to the addition of ligand, few peaks had bigger chemical shifts and some peaks slightly changed their position (Figure 8.28). The changing pattern of HSQC spectra due to the addition of L-glutamate had similarities with N-formyl glutamate, but it appeared that some peaks completely disappeared as a result of addition of L-glutamate which in case of N-formyl L-glutamate had slightly moved their position.

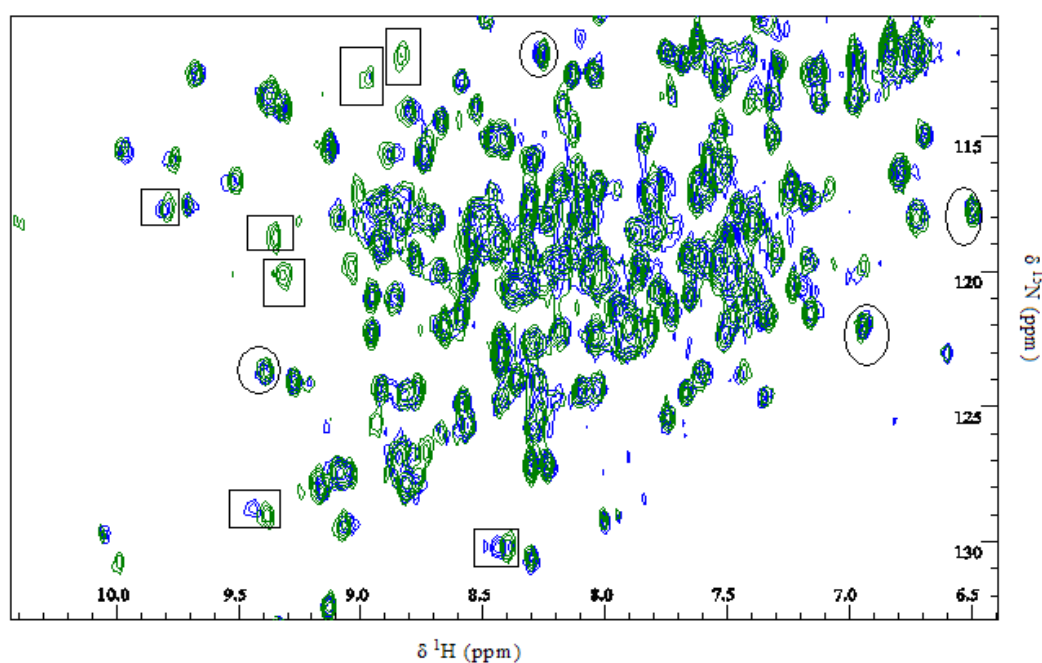
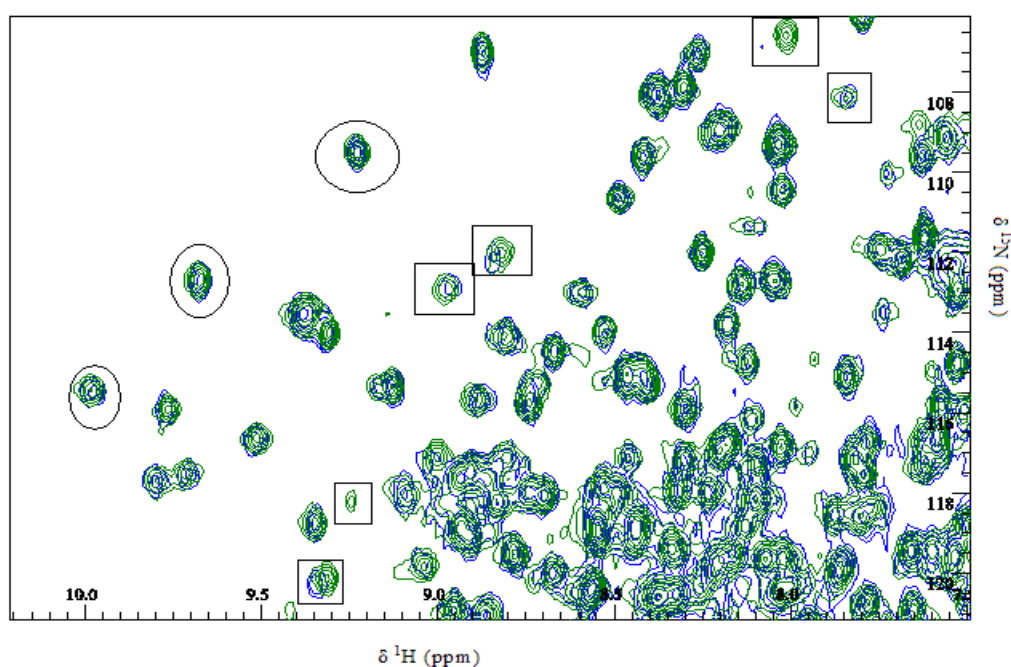


Figure 8.28 Overlaid HSQC spectra for PARI along with 20mM L-glutamate, empty square boxes around peaks represent change in position of peaks due to the addition of the ligand, empty circles around peaks represent no change due to the addition of ligand (green spectra = protein alone, blue spectra = protein with ligand)

### 8.7.4 Carbamoyl glutamate with PARI

In case of addition of carbamoyl glutamate to the protein, it was observed that some of the peaks moved slightly, while few peaks completely disappeared which may suggests the change in the conformation of some amino acids of the protein (Figure 8.29). By comparing the spectra of the protein alone and in the presence of carbamoyl glutamate, the spectral changes indicate that the chemical environment of some residues of the protein changed significantly which resulted in change in chemical shift values for certain amino acid residues.



**Figure 8.29** Overlaid HSQC spectra for PARI along with 20mM carbamoyl glutamate, empty square boxes around peaks represent change in position of peaks due to the addition of the ligand, empty circles around peaks represent no change due to the addition of ligand (green spectra = protein alone, blue spectra = protein with ligand)

### 8.7.5 Carbamoyl lysine with PARI

Both slight changes and bigger changes were observed in the overlaid HSQC spectra for carbamoyl lysine. The changing pattern of HSQC spectra due to the addition of carbamoyl lysine had similarities with carbamoyl histidine (Figure 8.30). The spectra were not clean to a desired level due to poor signal to noise ratio and thus caused difficulty in analyzing the spectral changes.

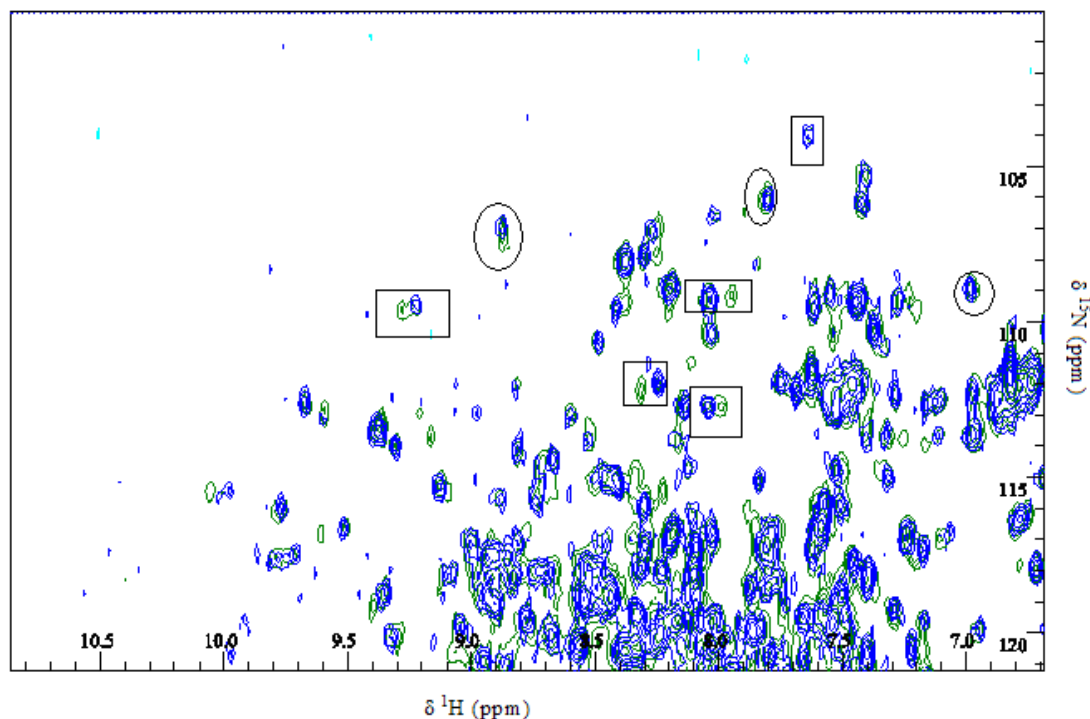
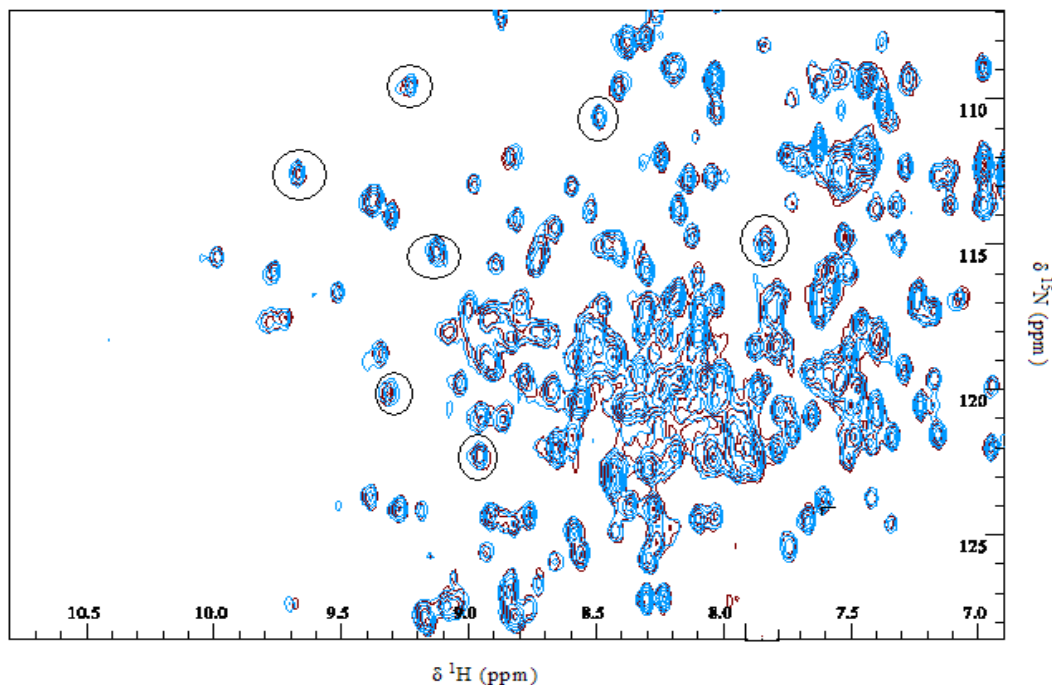


Figure 8.30 Overlaid HSQC spectra for PARI along with 20mM carbamoyl lysine, empty square boxes around peaks represent the change in position of peaks due to the addition of ligand, empty circles around peaks represent no change due to the addition of ligand (green spectra = protein alone, blue spectra = protein with ligand)

### 8.7.6 Carbamoyl cysteine with PARI

In the event of addition of carbamoyl cysteine to the protein very few changes were observed (Figure 8.31). By close observance it appeared that the spectrum for the protein with and without carbamoyl cysteine were almost similar. This indicates that carbamoyl cysteine does not bind to the protein at all.



**Figure 8.31** Overlaid HSQC spectra for PARI along with 20mM carbamoyl cysteine, empty square boxes around peaks represent the change in position of peaks due to the addition of ligand, empty circles around peaks represent no change due to the addition of ligand (green spectra = protein alone, brown spectra = protein with ligand)

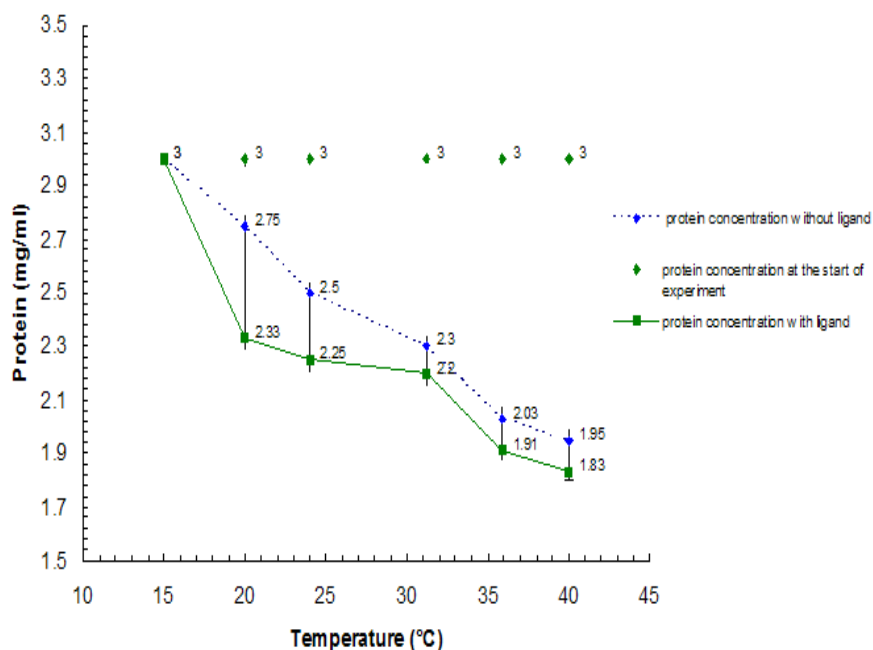
By analyzing all the individual and overlaid HSQC spectra it appeared that there were a number of phenomena taking place due to the addition of ligand to the protein. The slight change in chemical shift values of certain residues can be due to the change in the chemical environment of the residues. The disappearance of certain peaks as a result of ligand addition can be attributed to a complete conformational change of the amino acid residues. The bigger chemical shift changes in certain residues can be due to direct interaction with the ligand, which could be occurring in the protein active site. The increase/decrease in intensity of some peaks can be ascribed to more exposed/buried nature of the individual amino acid residue due to the addition of ligand. It was not possible to calculate the  $K_d$  value for individual ligands, which requires number of points in the form of ligand addition. The outcome of the experiments

was helpful in differentiating the ligand in terms of binders and non-binders. As explained by the HSQC the individual ligands can be categorized in terms of causing overall changes in descending order as:

**N-formyl L-glutamate > carbamoyl histidine > L-glutamate > carbamoyl glutamate = carbamoyl lysine > carbamoyl cysteine**

## 8.8 Effect of ligand addition on aggregation of PARI

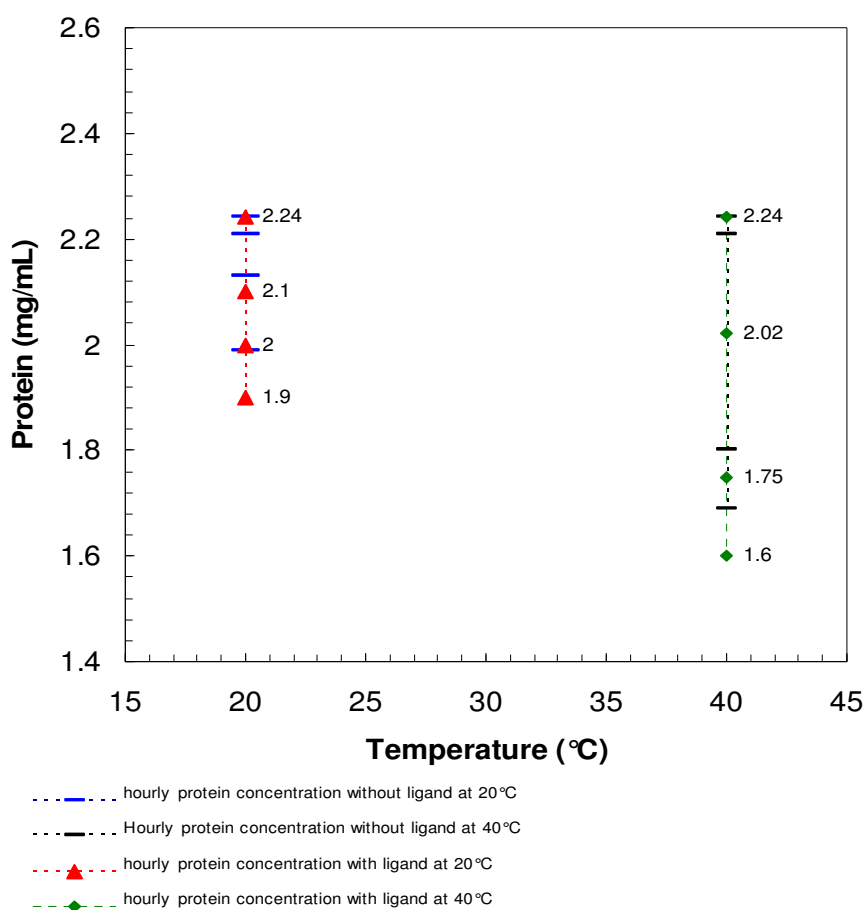
The addition of individual ligands to the protein caused aggregation which could be due to high temperature (35°C) and/or the addition of high concentration of ligand. During the NMR and CD experiments it appeared that the protein tends to precipitate quickly due to the addition of ligands. To find out the effect of ligand addition on the rate of aggregation a test experiment was carried out in the same manner as previously carried out for native protein (section 8.2). 20mM L-glutamate was added as a ligand to the individual protein samples prior to the start of the experiment. After a 40 minutes experiment it was observed that ligand addition significantly increases the aggregation rate, particularly in the 20–25°C and slightly in the 30–40°C temperature range (Figure 8.32).



**Figure 8.32** Graphical representation of effect of temperature on aggregation rate of PARI due to the addition of ligand at a range of temperature (vertical lines show the relative decrease in protein concentration due to the addition of ligand in a 40 minute experiment)



To further confirm these results another test experiment was carried out (as described in section 8.2) in which the protein samples along with the ligand were subjected to 20°C and 40°C temperatures for 3 hours and the protein concentration was measured after each hour (Figure 8.33). The comparative graphical data demonstrated that the liganded-protein tends to aggregate quickly and at higher rates than the unliganded-protein. It was concluded from the experimental data that the rate of aggregation is mainly due to the temperature factor, but the addition of ligand causes significant increase in it.



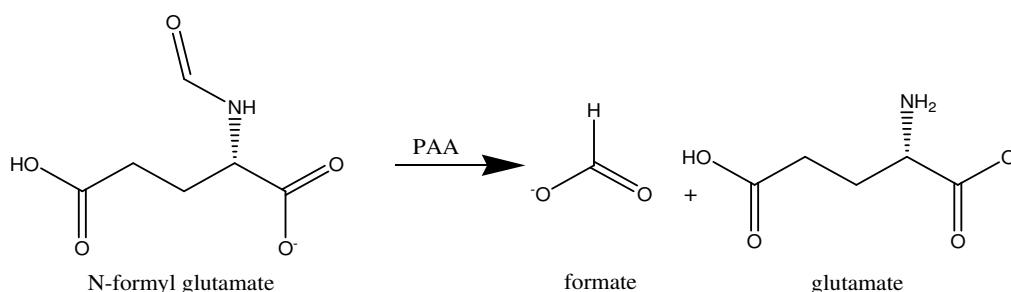
**Figure 8.33 Graphical representation of effect of temperature on aggregation rate of PARI (with and without ligand), the protein concentration was measured at 20°C and 40°C after each hour for a total of 3 hours**

## 8.9 Conclusions and Future work

The amount of work that could be performed on PARI was limited by significant issues which included low yield of protein from large scale cultures, prompt aggregation and instability of the protein due to the addition of ligand. By carrying out standard enzyme assays on PARI (PDB: 2Q7S) it is confirmed that it is not an NFGase. There is strong homology of the active site with Carboxypeptidase, not only the residues that coordinate the zinc atom but also catalytic residues and those associated with binding the c-terminal carboxylate group. Pharmacophore searching identified glutamate among others as possible residues to bind in the S1<sup>^</sup>. However in PARI structure there is no hydrophobic pocket S1, which would restrict the substrate specificity of the enzyme. It is presumed that upon peptide binding a conformational change accompanying the down ward movement of TYR227 side chain is necessary to open the S1 binding pocket. The NMR titrations suggest that peptides with a C-terminal glutamate may be potential substrates; however more binding experiments would be required to confirm this with other N-carbamoyl-amino acids to observe and define the specificity. An obvious next step is to synthesize dipeptides with a C-terminal glutamate and see whether they act as substrates by using the standard assay developed for testing carboxypeptidase activity. In the absence of peptidase activity, the assignment of this protein as a carboxypeptidase is uncertain.

## 9. Biochemical characterization of *Pseudomonas auroginosa* Amidohydrolase (PAA)

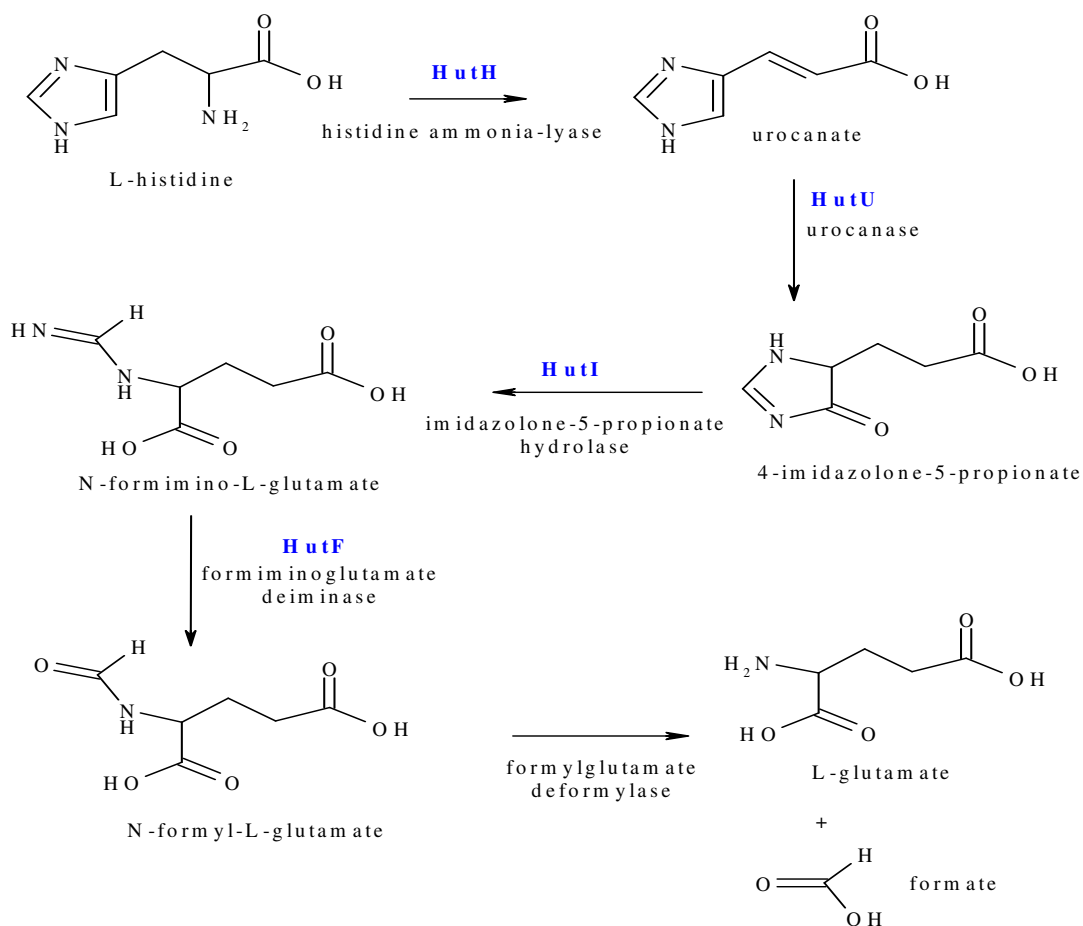
The N-formyl glutamate amido hydrolase from *pseudomonas auroginosa* (PAA) is the last enzyme in the Histidine utilization pathway II (Hut II). The enzyme hydrolyses N-formyl glutamate to glutamate and formate (Figure 9.1). The enzyme comprises of 266 amino acids with a Mr of 29.8Kda. In HUT II glutamate and formate are produced in the last step, but dependent on the organism it can produce formate or formamide. Formate is potentially toxic to the bacteria as it can form formaldehyde and can be simply metabolized to CO<sub>2</sub> by formate dehydrogenase {165}. Interestingly majority of the formate from this pathway gets incorporated as the C2 of purine rings in purine biosynthesis {166}. The structural mechanism by which formate is incorporated as C2 in the purine rings during purine biosynthesis is unknown.



**Figure 9.1 PAA catalyzing N-formyl glutamate to formate and glutamate**

Hut I comprises of 4 enzymatic steps while the Hut II pathway has 5 enzymatic steps. The initial 3 enzymatic steps are same in both Hut I and Hut II. The two pathways differ in their conversion of N-formimino-L-glutamate (FIGLU) to L-glutamate. In Hut I pathway the conversion is achieved by a single enzymatic step involving formiminoglutamate hydrolase (HutG), which hydrolyses FIGLU to yield L-glutamate and formamide. In Hut II pathway the same overall process is achieved by 2 enzymatic reactions where FIGLU is initially converted to N-formyl-L-glutamate (NFGLU) by formiminoglutamate deiminase (HutF) and in the

final 5<sup>th</sup> step NFGLU is hydrolyzed by formylglutamate deformylase (PAA) to yield L-glutamate and formate (Figure 9.2).



**Figure 9.2** Hut II 5 step pathway, PAA catalyzes the 5<sup>th</sup> step by converting NFGLU to L-glutamate and formate, image adopted from {151}.

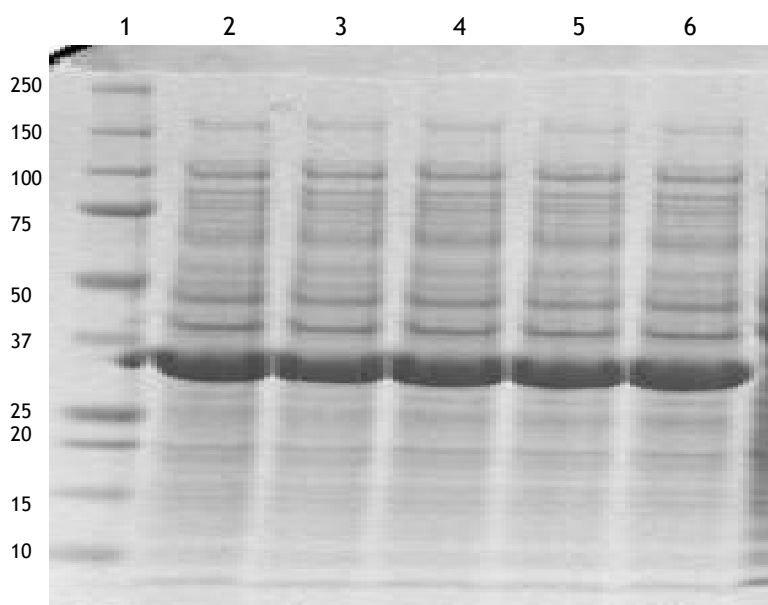
## 9.1 Aims and objectives

The aims of the study which could help in understanding the enzyme mechanism are given below

1. Kinetic characterization of the enzyme
2. Inhibition assays on the protein with potential inhibitors mimicking the true substrate of the protein
3. Co-crystallization of protein with different inhibitors for analysis of binding mode.

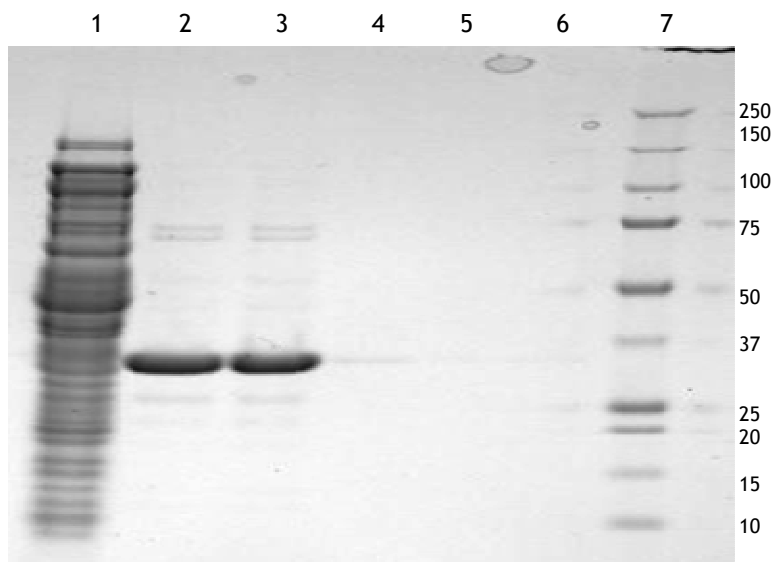
## 9.2 Expression and purification of PAA

The plasmid was transformed in to BL21 (DE3) cells, and the protein was over expressed in auto induction media (Figure 9.3). The N-terminal His tagged protein was then purified through Nickel column.



**Figure 9.3 SDS-PAGE analysis for PAA after Over night induction in autoinduction media: 1= molecular weight marker in Kda (Precision plus protein<sup>TM</sup> BioRad, cat# 161-0373), 2-6= overnight induced samples of PAA (vertical column with different molecular weights for individual bands)**

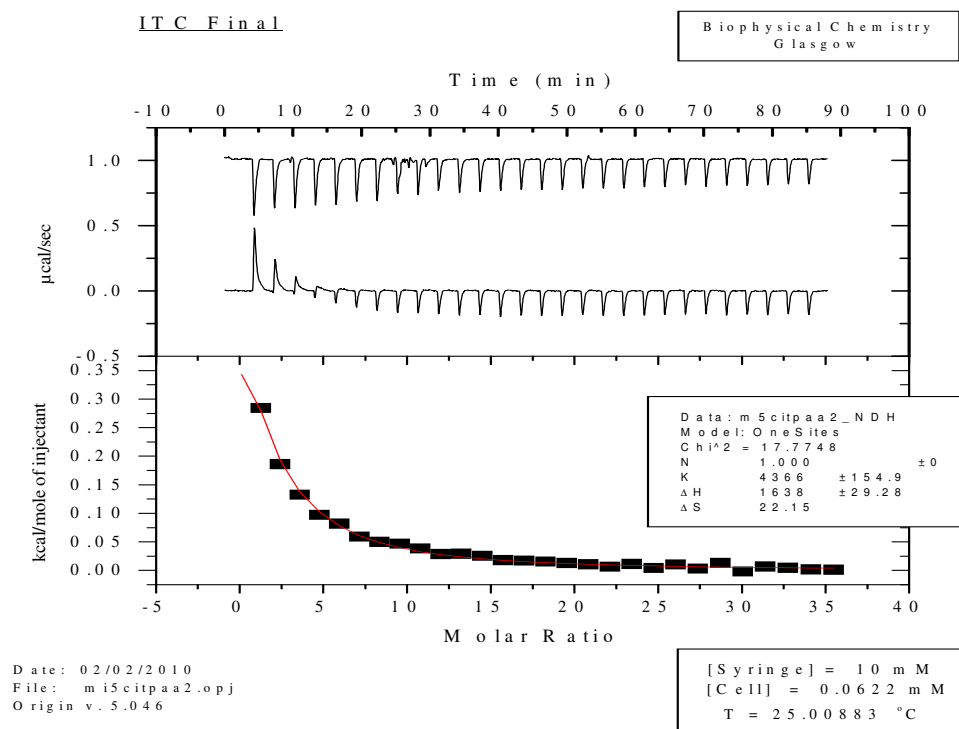
SDS-PAGE analysis showed that the protein is eluted at 75mM Imidazole concentration instead of 200 or 300 mM Imidazole concentration, keeping this in view in all the future batches the PAA were purified by eluting at 75mM Imidazole with a wash at 20mM Imidazole (Figure 9.4). For all the SDS-PAGE analysis NuPAGE Novex (Invitrogen®) Bis-Tris 4-12 % gel was used in 4-12 % Bis-Tris (MES) buffer. The elution fraction was buffer exchanged to get rid of excess of Imidazole and subsequently concentrated down to the desired concentration by using Vivaspin®. The concentrated protein was then used for co-crystallizations, Isothermal titration calorimetry (ITC) and inhibition assays.



**Figure 9.4 SDS-PAGE analysis for PAA after Ni-purification:** 1= flow through, 2= 75mM Wash1, 3=75mM Imidazole wash2, 4, 5= 200mM Imidazole elute, 6= 300mM Imidazole elute, 7=molecular weight marker in Kda

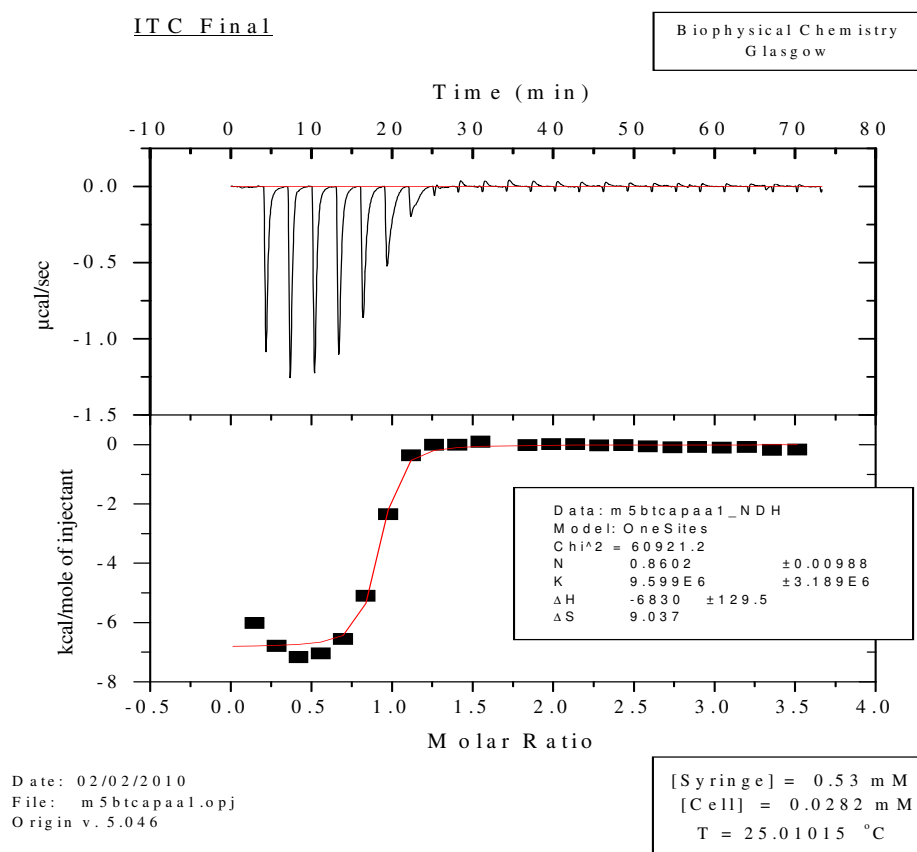
### 9.3 I T C studies on PAA with potential inhibitors

ITC experiments were carried out on protein, with the potential inhibitors, butane tri carboxylic acid (BTCA) and Citrate. 1mM of citrate was used in the beginning but the titration curve did not show any change, therefore later on the concentration was increased to 10mM. The increase in concentration resulted in an observable change in the titration curve (Figure 9.5).



**Figure 9.5** ITC titration Curve for PAA along with citrate

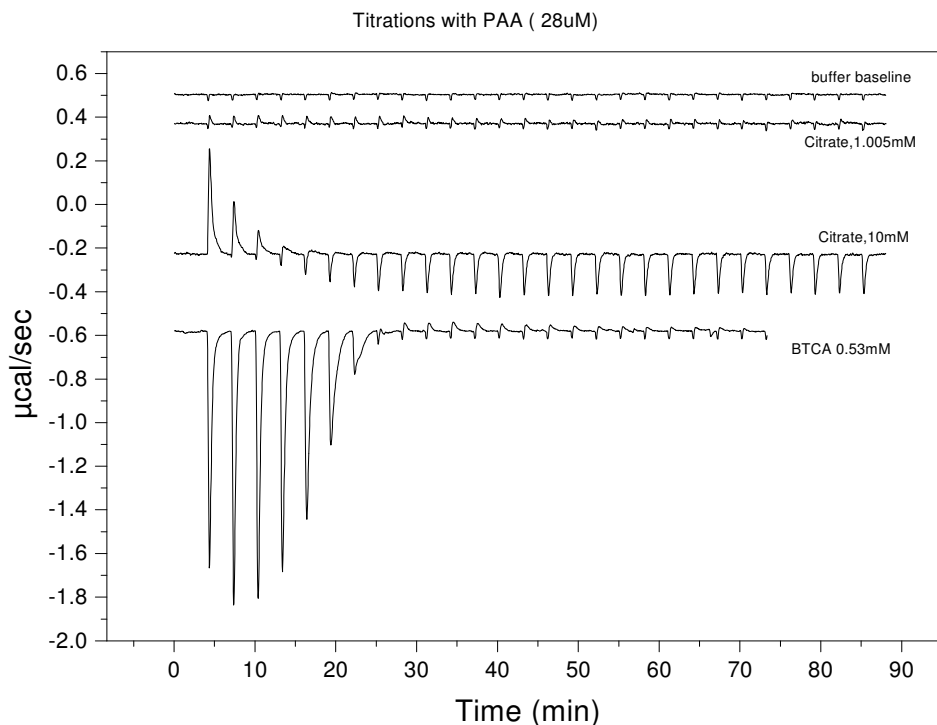
For BTCA comparatively tight binding was observed as the initial 0.53mM concentration was sufficient to show changes in the titration curve (Figure 9.6). The ITC data obtained after titration gave N value of 0.86 for BTCA indicating that inhibitor and binding site ratio is about 1:1, K value (Association constant,  $K_a$ ) was 9.6 $\mu$ M, which in terms of  $K_d(1/K_a)$ = 0.1 $\mu$ M, the enthalpy value ( $\Delta H$ ) was negative showing that binding of BTCA to PAA is exothermic in nature.



**Figure 9.6 ITC titration Curve for PAA along with BTCA**

The binding curve for citrate gave N value of 1, showing 1:1 binding of a citrate molecule to PAA, the K value was 4366M ( $K_d = 0.23\text{mM}$ ), enthalpy value ( $\Delta H$ ) was positive (1638), showing that the reaction is endothermic in nature, and the entropy value ( $\Delta S$ ) was fairly high (22.15) in comparison to BTCA (9.0), showing that the reaction between citrate and PAA is more entropy driven than BTCA with PAA. ITC results were further supported by the enzyme inhibition assays, which gave  $K_i$  values of 150nM and 0.54mM for BTCA and citrate respectively.



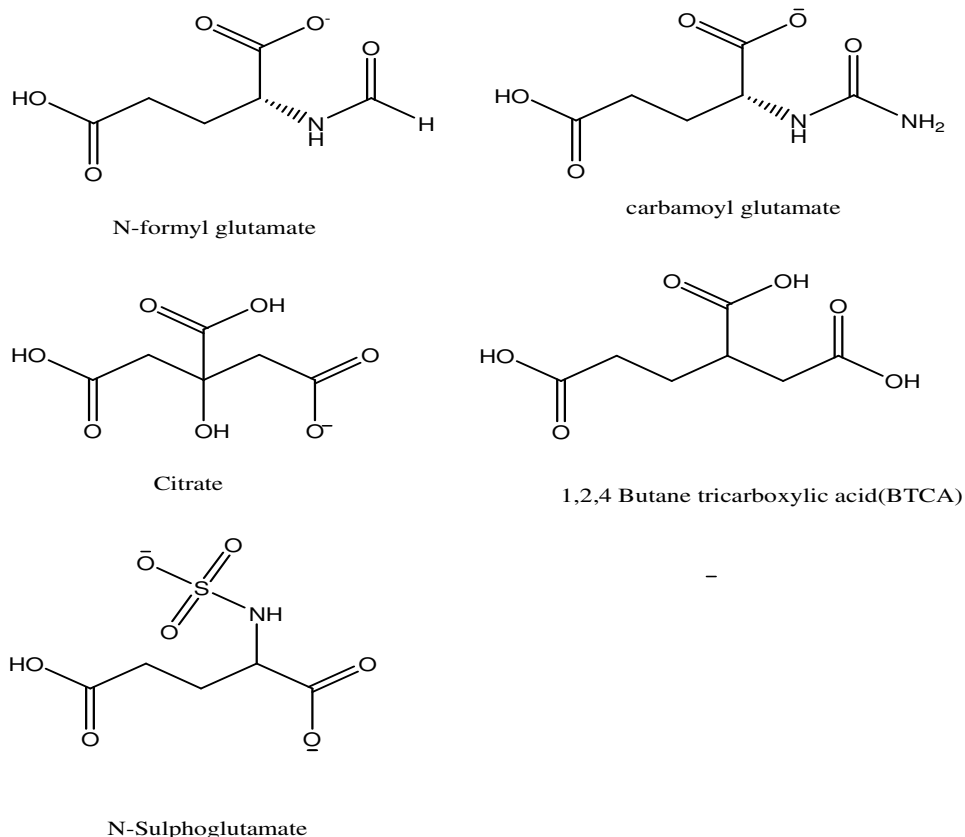


**Figure 9.7 Overlaid ITC titration curve for PAA with both BTCA and Citrate**

The ITC results demonstrated tight binding of PAA to BTCA while weak binding to citrate. The concentration of protein used in the ITC experiments was 28μM against Citrate (10mM) and 62μM against BTCA (0.53mM).

## 9.4 Inhibition studies on PAA

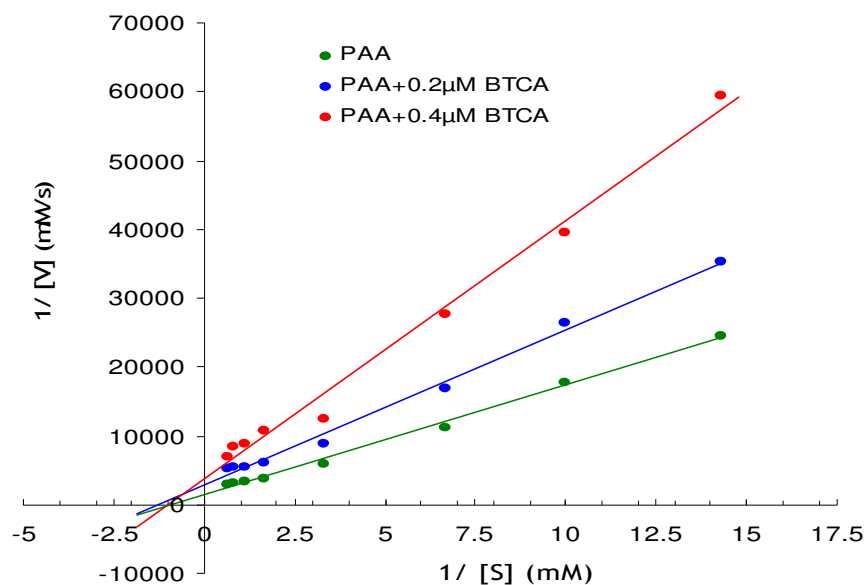
Previously enzyme assays have been carried out for PAA in the presence and absence of inhibitors (Teresa Morenés Bertrán an Erasmus student). Inhibitors which have been tested previously include EDTA (non competitive inhibitor  $K_i = 110\text{nM}$ ), Citrate (competitive inhibitor,  $K_i = 0.54\text{mM}$ ) and potential transition state mimic N-Sulphoglutamate (competitive inhibitor,  $K_i = 2.36\text{mM}$ ). The potential inhibitor 1, 2, 4- Butane tricarboxylic acid (BTCA) and carbamoyl glutamate (CG) were selected on the basis of structural similarity with citrate and the substrate N-formyl glutamate (Figure 9.8).



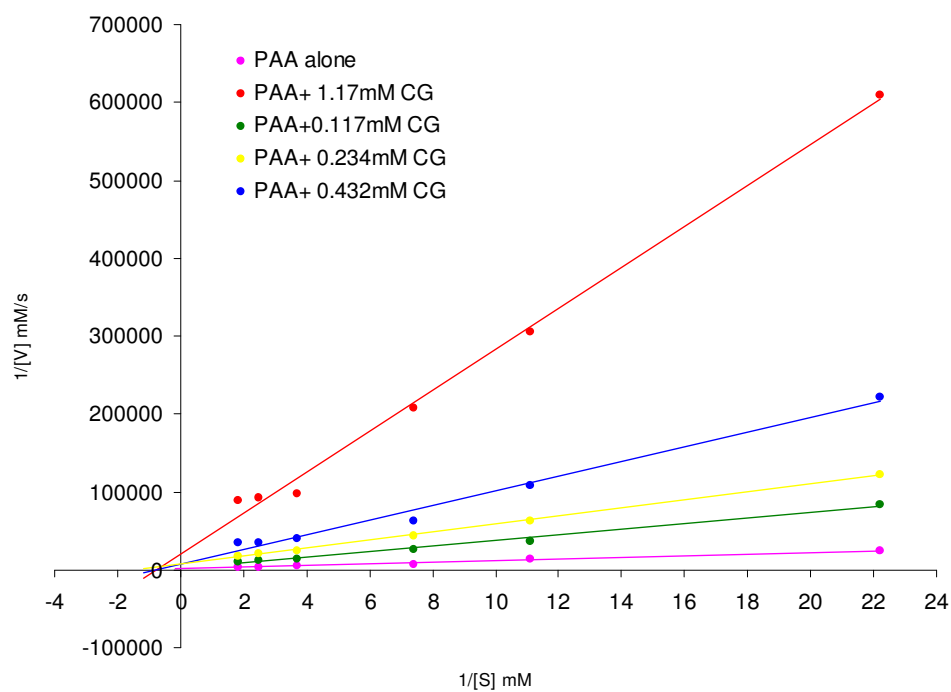
**Figure 9.8 Comparison of chemical structures of substrate (N-formyl glutamate) with inhibitors (BTCA, Citrate and CG)**

The enzyme assays were performed by using jasco-550 double beam spectrophotometer at a wavelength of 210nm. The hydrolysis of amide bond was followed by a decrease in absorbance at 210nm. Both enzyme and inhibition assays were carried out in PBS buffer (1.4M NaCl, 27mM KCl, 100mM Na<sub>2</sub>HPO<sub>4</sub>, 18mM KH<sub>2</sub>PO<sub>4</sub> pH: 7.2). Final concentration of protein was 1.38μM and 0.46μM in the cell for BTCA and CG respectively. The substrate concentration was varied between 0.05–1.5mM. Different inhibitor concentrations were used for calculating the K<sub>i</sub> values. Inhibition assays on both BTCA and CG indicated non competitive inhibition with a K<sub>i</sub> of 150nM (similar to EDTA) for BTCA and 40μM for CG (Figure 9.9-10). The Lineweaver- Burk equation given below was used for calculating the K<sub>m</sub>, V<sub>max</sub> and slope (K<sub>m</sub>/V<sub>max</sub>) values.

$$\frac{1}{V} = \left( \frac{K_m}{V_{max} [S]} + \frac{1}{V_{max}} \right)$$



**Figure 9.9** Line weaver-Burk plot for PAA with and without BTCA (the extrapolated lines cross at the same point at negative x-axis to give the same  $-1/K_m$  value)



**Figure 9.10** Lineweaver-Burk plot of PAA with and without CG (the extrapolated lines cross at the same point at negative x-axis to give the same  $-1/K_m$  value)

### 9.4.1 Calculation of $K_i$ for BTCA and CG

The inhibition constant,  $K_i$  was calculated by using the appropriate formula for noncompetitive inhibition as mentioned in figure 9.11

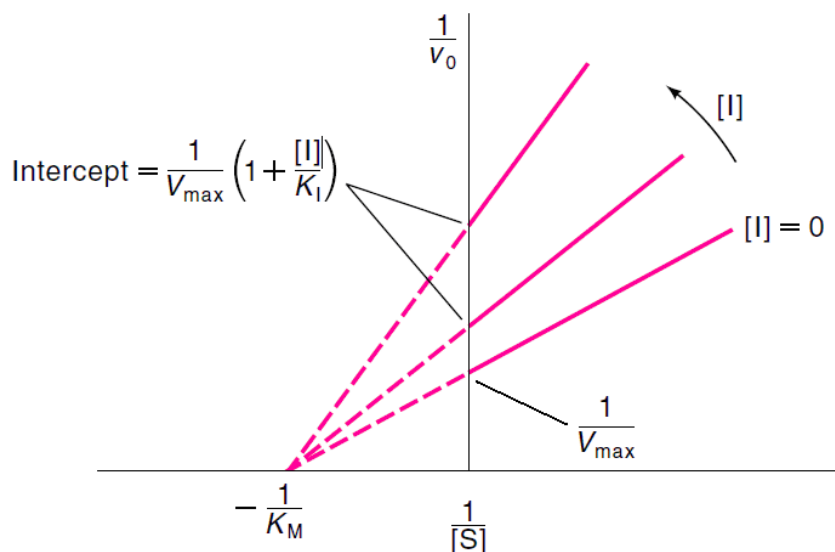


Figure 9.11 Graphical representation of Lineweaver-Burk plot for Non-competitive inhibition

The  $K_i$  values calculated for BTCA and CG at individual inhibitor concentrations are given in table 9.1. The average calculated  $K_i$  value for BTCA was 190nM, which is near to the  $K_i$  value for EDTA (117nM, previously determined in the group). The average  $K_i$  value for CG was calculated to be 34 $\mu$ M.

Inhibitor	Inhibitor concentration (mM)	Calculated $K_i$ value ( $\mu$ M)
BTCA	0.2	0.240
	0.4	0.140
CG	0.117	53.6
	0.234	40
	0.468	32
	1.17	40

Table 9.1  $K_i$  values for PAA against BTCA and CG by using non-linear regression plots.

The  $K_m$  and  $V_{max}$  values for BTCA and CG were calculated by using both Lineweaver-Burk (Table 9.2) and non linear regression plots (Table 9.3).

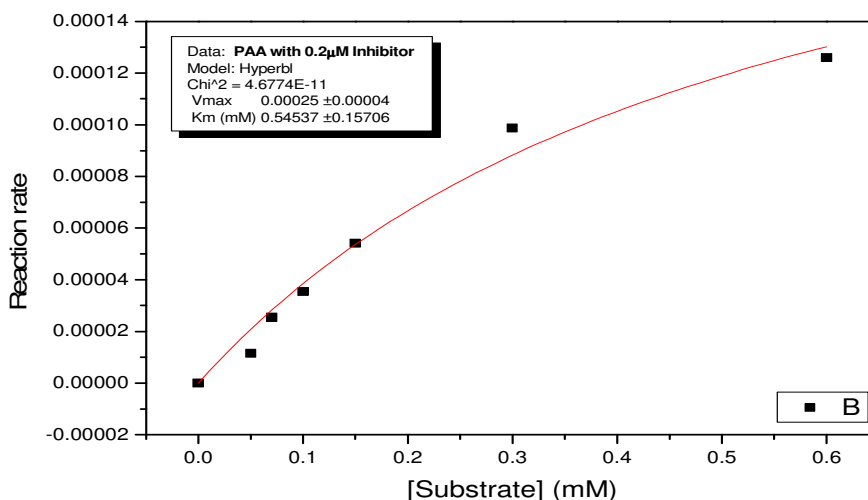
Inhibitor	Inhibitor concentration	$K_m$ (mM )	$V_{max}$ ( $\mu$ M/S )
BTCA	0.00 $\mu$ M	1.19	12.3
	0.2 $\mu$ M	0.80	5.90
	0.4 $\mu$ M	0.7611	3.48
CG	0.00 mM	0.776	4.34
	0.117 mM	0.784	2.53
	0.234 mM	1.31	2.32
	0.468 mM	1.37	1.28
	1.17 mM	2.12	1.32

**Table 9.2  $K_m$  and  $V_{max}$  Values of PAA with BTCA and CG by using Lineweaver-Burk plots.**

The values obtained through Lineweaver-Burk plot are not the most accurate way of obtaining these values as the plot and the trend line are overly affected by points measured for the lowest substrate concentrations, which have the largest error. A better method is to calculate the values directly by non linear regression fit to the Michelis-Menton equation (Table 9.3).

Inhibitor	Inhibitor concentration	$K_m$ (mM )	$V_{max}$ ( $\mu$ M/S )
BTCA	0.00 $\mu$ M	0.6124 $\pm$ 0.0563	15.33 $\pm$ 3.81
	0.2 $\mu$ M	0.5454 $\pm$ 0.1571	8.33 $\pm$ 1.33
	0.4 $\mu$ M	0.3346 $\pm$ 0.0719	4.0 $\pm$ 0.26
CG	0.00 mM	0.692 $\pm$ 0.158	12.4 $\pm$ 2.68
	0.117 mM	0.687 $\pm$ 0.180	3.9 $\pm$ 0.65
	0.234 mM	0.504 $\pm$ 0.05	1.84 $\pm$ 0.19
	0.468 mM	0.288 $\pm$ 0.06	0.79 $\pm$ 0.05
	1.17 mM	0.441 $\pm$ 0.23	0.40 $\pm$ 0.09

**Table 9.3  $K_m$  and  $V_{max}$  Values of PAA with BTCA and CG obtained by using Non-linear regression plots.**



**Figure 9.12** A typical Non-linear Regression plot for PAA with 0.2  $\mu\text{M}$  BTCA ( $V_{\text{max}}$  value is in M/min)

For BTCA and CG the individual kinetic inhibition data in terms of Lineweaver-Burk plot gave the  $-1/K_m$  values which were close to each other. Very less variation in  $-1/K_m$  value both in the inhibited and un-inhibited cases indicated a Non-competitive inhibition against PAA. The analysis of data from both the plots indicated greater changes in  $V_{\text{max}}$  values in comparison to the  $K_m$  values which again assured that the inhibition is non-competitive in nature.

#### **9.4.2 PAA Inhibition studies using Nickel, EDTA, BTCA and Zn solution**

To test if active site zinc is being chelated by BTCA in the same way as EDTA. The inhibition assays were performed in the presence of  $\text{Ni}^{2+}$  solution. In case of a chelating effect BTCA would interact with  $\text{Ni}^{2+}$  preferentially and therefore would result in reduced inhibition. The inhibition assays were carried out at 0.2  $\mu\text{M}$  BTCA with 1mM solution of  $\text{NiCl}_2$  in the reaction mixture. No change in the rate of reaction was observed in comparison to the control experiment (0.2  $\mu\text{M}$  EDTA along with nickel solution). The rate of reaction was faster when EDTA was used along with the Nickel solution and was half when only EDTA was used

(Table 9.4). The inhibition assays thus confirmed that BTCA is not chelating the zinc in the active site.

S.No	Amount of Enzyme( $\mu$ M)	Substrate Conc. (mM)	EDTA Concentration	NiCl <sub>2</sub> concentration	BTCA concentration	$\Delta$ Abs/min
1	1	0.6	-	-	-	0.728
2	1	0.6	-	-	0.2 $\mu$ M	0.444
3	1	0.6	-	1mM	-	0.748
4	1	0.6	-	1mM + (Pre-incubated)	0.2 $\mu$ M	0.473
5	1	0.6	-	-	-	0.730
6	1	0.6	0.2 $\mu$ M	-	-	0.340
7	1	0.6	0.2 $\mu$ M	1mM	-	0.681

**Table 9.4 Enzyme kinetics for PAA in the presence of Ni<sup>2+</sup>, BTCA and EDTA (- = Not added).**

### 9.4.3 Dialysis of PAA with BTCA and Inhibition Assays

Non competitive inhibition of the enzyme could be due to the binding of the inhibitor at a site away from the substrate binding site. Alternatively, it could be due to the inhibitor binding to the active site covalently. Initially it was presumed that BTCA is acting in a similar way as EDTA by chelating zinc in the active site. This hypothesis was tested by adding inhibitor to the protein followed by dialysis and subsequent assays.

Status	Inhibitor concentration ( $\mu$ M)	Enzyme concentration ( $\mu$ M)	Substrate Concentration (mM)	$\Delta$ Abs/min
Before dialysis	0.00 $\mu$ M	1	0.6	0.721
	10.00 $\mu$ M	1	0.6	0.0080
After dialysis	0.00 $\mu$ M	1	0.6	0.713
	10.0 $\mu$ M	1	0.6	0.1460

**Table 9.5 Enzyme inhibition assays for PAA before and after dialysis**

Inhibition assays were carried out to calculate the saturating concentration of BTCA against PAA. The result showed that 10uM of BTCA was sufficient to completely inhibit 1uM of the enzyme. Two (native PAA and PAA with 2mM BTCA) 100uL samples of 0.3mM enzyme were dialyzed overnight at 4°C in 10mM sodium phosphate buffer pH:7.2. After 18 hours of dialysis, enzyme assays were carried out on both the samples (Table 9.5). The reaction rate of the post-dialyzed enzyme incubated with BTCA was 20 times more than the pre-dialyzed enzyme incubated with BTCA. The regain in activity observed could be only if the inhibitor has diffused out of the enzyme during dialysis. The significant recovery in activity of PAA after dialysis lead to the conclusion that BTCA is neither an irreversible inhibitor which chelates the zinc nor it is forming an irreversible covalent link to the enzyme.

## 9.5 Crystallization trials for PAA with BTCA

The Automated crystallization trials were carried out for PAA with Inhibitor 1, 2, 4 Butane Tri carboxylic acid (BTCA). The crystallizations were carried out at a final concentration of 0.7mM (20mg/mL) of protein and 4mM concentration of inhibitor. The crystallizations were carried out by the robot system using the PACT and the JCSG+ suite composition, the M-Screen (M1-M72) was also tried manually, crystalline precipitates were observed in M-Screen including M-05, M-09, M-10, M-20, M-26 and M-72 . In case of PACT ½ screen crystalline type of precipitates were seen in E4, E8, E9 and H11 in both native protein and with inhibitor. Keeping in view the PACT ½ conditions at which the crystalline precipitates were obtained, crystallizations conditions were further optimized. Crystalline precipitates were seen at PEG 8K concentration of 4%, 6%, 8% and 10%, Increase nucleation was causing a shower of small crystals and Nucleation rate seemed to be very high due to which the individual crystals were unable to grow. Streak seeding was also carried out by using a cat hair in to a clear drop to cause the seeding. The streak seeding resulted in crystalline precipitates but not individual crystals suitable for diffraction studies.



## 9.6 Conclusions

The following conclusions can be drawn from the enzyme inhibition studies.

1. BTCA is a tight binding noncompetitive inhibitor of PAA. Although the inhibitor has a similar  $K_i$  to that of EDTA, but it is experimentally demonstrated that its effect is not by chelation (as by EDTA) of the active site  $Zn^{2+}$ , but rather by binding to the enzyme directly.
2. In case of BTCA, the nature of the inhibition has been investigated and indicates that the inhibitor can be dialyzed out of the enzyme over time and therefore does not form an irreversible covalent adduct with the protein.

## 9.7 Future Aims

In order to characterize the mechanism of enzyme, it would be appropriate to obtain co-crystals of the enzyme with the inhibitors described here. Methods to obtain crystals include changing the protein source from related bacteria, and limited proteolysis, for example. A crystal structure would greatly help in understanding the substrate specificity of the enzyme and give a structural context to the data presented here.

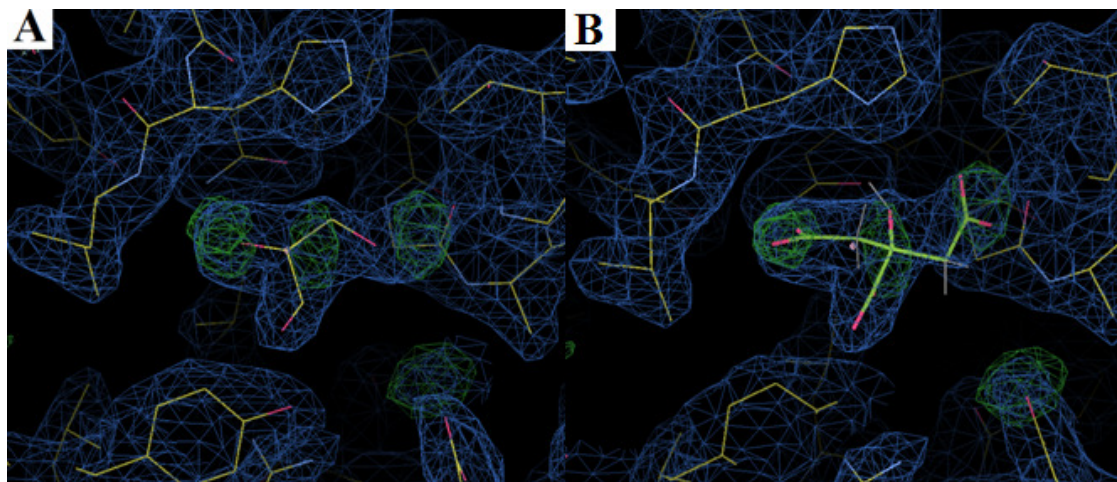
## 10. Probing protein X-ray structures with unknown ligands

### 10.1 Introduction

The very nature of the protein data bank (PDB) is the possibility to deposit an X-ray structure of protein with an unknown ligand or a completely wrong ligand. This may be due to an eagerness to publish and secondly perhaps less rigorous supervision, which finally may result in errors in the submitted X-ray structure. Mostly the errors related to the correct assignment of ligands and co-factors are because they may be of little importance to the publishing author. For instance a structural biologist may be interested more in the overall fold of the protein, but becomes a nightmare for a medicinal chemist who may be interested very specifically in the binding site {167}.

In the past majority of the protein X-ray structures were deposited with the aims and interest towards understanding the overall fold of the given protein, and thus the bound ligands or co-factors were not given the same consideration as the protein backbone. There were program libraries available for fitting the residues in geometrical reasonable conformations, but none for the ligands and co-factors, as a result they refined into unlikely geometries. In recent years due to a huge increase in the use of protein crystallography in structure based drug design, the protein-ligand interface has been recognized as important as the overall fold of the protein, and more attention is being paid now in order to assure that the ligands and co-factors are treated appropriately {167}.

Some authors have recently pointed out that some of the deposited structures of proteins containing ligands in the PDB database are totally wrong. Some of these structures may be wrong due to a mistake by a crystallographer or because of misunderstanding of the chemistry that is modeled, perhaps by introducing some error in the stereochemistry {168-170}. For instance figure 10.1 shows the typical assignment of a wrong ligand in the PDB structure of 2UYG, where the electron density fits by a citrate molecule but the structure has been deposited with a glycerol molecule as a ligand.



**Figure 10.1** Wrong identification of a ligand as (A) a glycerol molecule instead of (B) a citrate molecule in X-ray structure of 2UYG, the blue electron density is the weighted observed electron density and the green density is difference density showing missing components of the ligand.

Apart from human errors there are two other reasons for poor ligand geometry. First one is the resolution of diffraction pattern obtained from protein structures which are so bad that it becomes very difficult to refine them for individual atom positions. Second error culminates due to the insufficient search of various other viable ligand conformations during the model building phase {167}.

To address these issues and to propose a reasonable solution, Discovery studio visualizer (DSV) and Catalyst® were used in this part of the project. The programs were used for pharmacophore model development and then searching against the databases. The databases used were naturalism, comprising of 5,422 small chemicals created from the ChEBI database, or the full database of 14,000 compounds. The X-ray structures with unknown ligand(s) were identified with code UNL, selected through the RSCB web interface.

## 10.2 Aims and Objectives

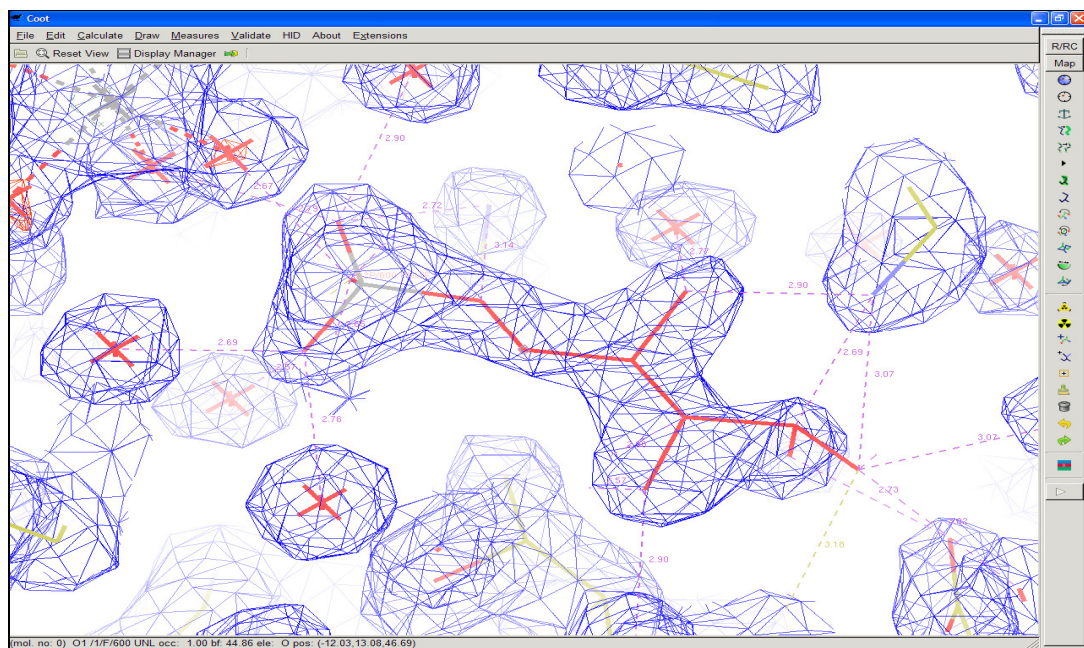
The study was carried out with the following aims:

1. If pharmacophore searching can be utilized to conceive the identity of unknown ligands.

2. What are the limitations due to low resolution data in correct identification of ligands?
3. How well the pharmacophore hits satisfy the electron density of the unknown ligands?

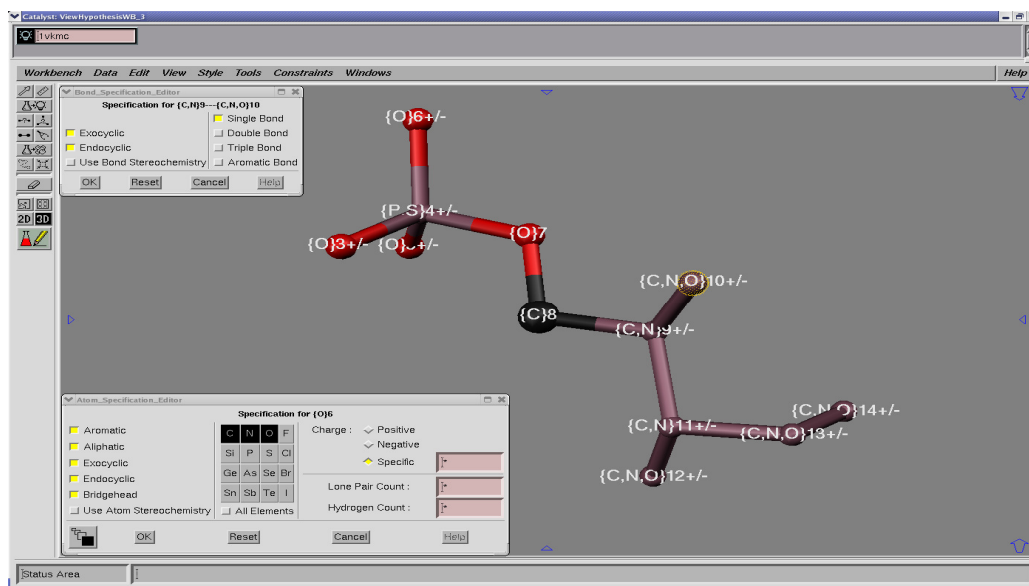
### 10.3 Screening of Protein X-ray structures

The RCSB protein data bank (PDB) ([www.rcsb.org](http://www.rcsb.org)) was searched for the occurrence of unknown ligands as a starting set of structures to evaluate. It is essential to have electron density of the unknown ligands to experimentally evaluate them. The coordinates and data were uploaded from the Uppsala electron density server (EDS) {171} and visualized by using the graphics software Coot version 0.6.2 {95}. Initially all the available protein structures with unknown ligands in RCSB protein data bank were visualized by using Coot. This allowed visual inspection of the electron density to confirm that the unknown ligand is in fact a proper ligand and not just a collection of water molecules or a polyethylene glycol (PEG) fragment or a single point presumed a metal ion. It was clear that there was a real variety of density with some researches labeling any non spherical density as unknown ligand, while others represented whole ligands in this way. These whole ligands which were the exception were chosen for consideration for ligand searching. By studying these unknown ligands in the protein binding site, in some cases the ligands were similar to certain compounds i.e. Glycerol, phosphate and citrate and it appeared that some of the compounds were co-crystallized with the protein because of their presence in the crystallization conditions. The X-ray liganded structures were downloaded from the PDB web link ([www.rcsb.org](http://www.rcsb.org)). All the X-ray structures were first visualized in Coot program (Figure 10.2) to observe and compare the calculated and observed electron density map (for the protein structure in general and for the unknown ligand in particular).



**Figure 10.2** Electron density map visualization in coot, unknown ligand atoms with observed electron density (blue mesh) around it in the binding site of the protein (1VKM), the dashed lines represent the inter atomic distances between the atoms of the unknown ligand and surrounding amino acid residues and water molecules

In Catalyst certain features of the pharmacophore model were parameterized by optimizing chemical features of the atoms of the unknown ligand i.e. atom type, charge on atom, bond type and hydrogen count on certain atoms (Figure 10.3). The bond order clearly gave some indication of a potential atom type but it was important not to try and first guess atom types. Hydrogen bond interactions with the protein gave further restraints to the query.



**Figure 10.3** Catalyst software with provision of optimizing bond type, multiple atomicity and hydrogen count on the atoms of the unknown ligand for final modelling of the pharmacophore (exclusion sphere have been hidden for clarity viewing)

Further details like information about the binding domain, location of unknown ligand (either present in the binding site or at the periphery of the protein), electron density, selected pharmacophore searches and best hits are represented in tabulated form in Appendix 1. During visualization of protein structures with unknown ligands it appeared that

1. Most of the structures had poor and disordered electron density.
2. In some cases the bonding electron density between the ligand atoms was absent.
3. Some ligands were lying in a complete hydrophobic pocket of the protein with the possibility of very few interactions.

Further details relating to electron density and other factors about the protein structures in numerals are given below in table 10.1

Over all details for PDB structures	Number of PDB structures
Structures with ligands having poor and/or disordered electron density for the unknown	70
Structures with ligand having very weak electron density	53
Structures without X-ray data	33
structures with disordered and fragmented ligands	19
Structures with benzamidine like ligand in the binding site	13
Structures with electron density absent between the atoms of the ligands	12
Structures with ligands located outside the binding pocket at the periphery of the protein	10
Structures with ligands in hydrophobic pocket of the protein	3
Structures subjected to pharmacophore searching	14
Total number of PDB structures evaluated	227

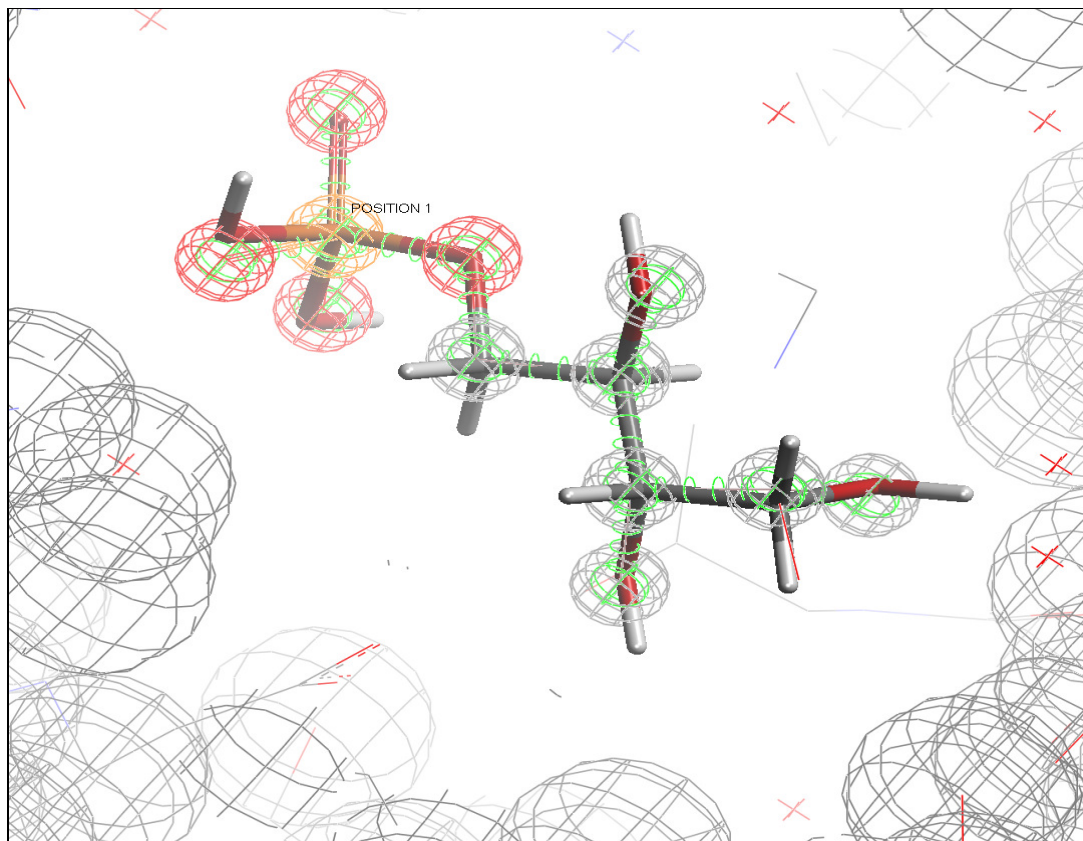
**Table 10.1 Total number of PDB structures and different attributes of the protein structure with respect to the ligand in separate rows with numerals**

The X-ray structures (with more well defined electron density of the unknown ligand matching with the position of atoms within in the binding site) were selected for pharmacophore searching. DSV was used for generating the pharmacophore model for the selected X-ray structures of the protein by using the DSV Vector and Query atom methods, as described in section 3.5.2-3. A number of factors were taken in to account for the selection of protein structures with unknown ligands for pharmacophore searching i.e. availability of potential H-bond interactions between the ligand and the protein, availability of good electron density for the unknown ligand and location of unknown ligand in well defined binding site. The pharmacophore searches which gave interesting results in the form of potential hits are mentioned below, while some searches which were not very conclusive are mentioned in the Appendix 1. Details of X-ray structures, their pharmacophore modeling and resulting hits conforming to the electron density of the unknown ligands are presented below under the PDB code title heading.



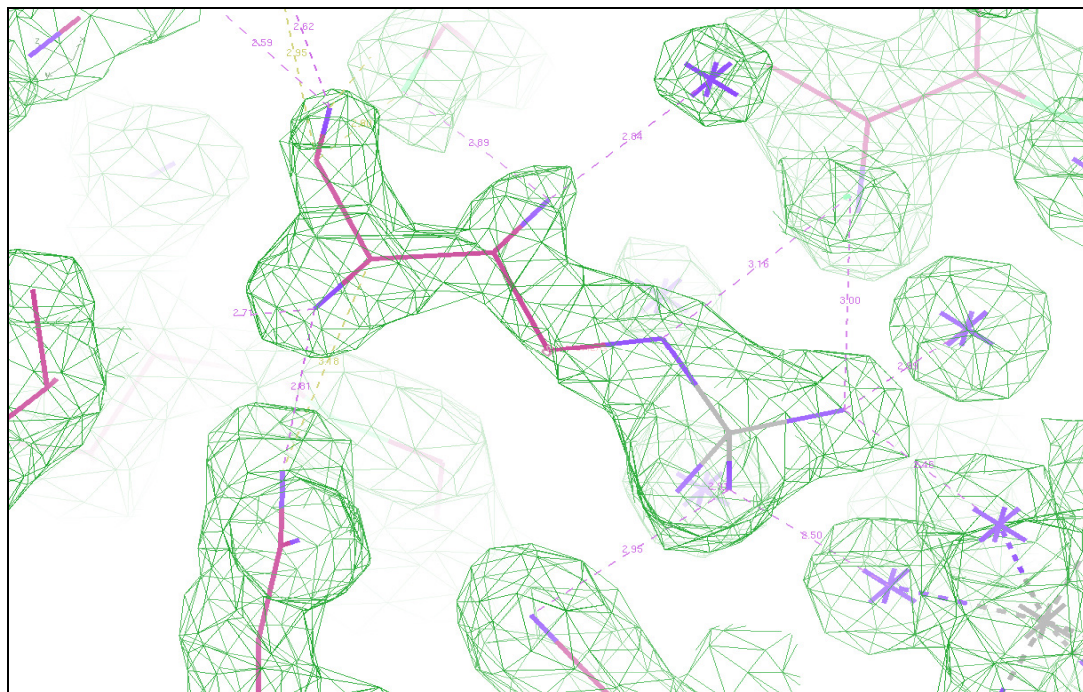
### 10.3.1 1VKM

1VKM is the Crystal structure of an indigoidine synthase a (idga)-like protein (tm1464) from *Thermotoga maritima* MSB8 diffracted at a resolution of 1.90Å. The hits obtained as a result of pharmacophore search model by using the query atom method, were fitting well in the binding site. A total of 29 hits were obtained from the final pharmacophore search. The majority of the hits were phosphate sugars including pentoses, hexoses and some 3C and 4C sugars. Assigning multiple atomicities to the atoms resulted in reasonable hits, especially in case of phosphate groups, the inclusion of phosphorus atom among other atoms (C, N) of the query atom at position 1 helped in finding various sugar phosphates, which satisfied the pharmacophore model (Figure 10.4). Among the hits D-erythritol 4-phosphate coherently fitted the e-density (Figure 10.5) and seems a unique hit which further needs to be investigated in connection with the mechanism of binding and any activity if present.



**Figure 10.4** D-erythritol 4-phosphate shown as stick model, efficiently satisfies all the constraints of the pharmacophore model, while few atoms of some other hits (shown as line model) violate the pharmacophore model (some of the amino acid residues and exclusion spheres are removed for clarity purposes)

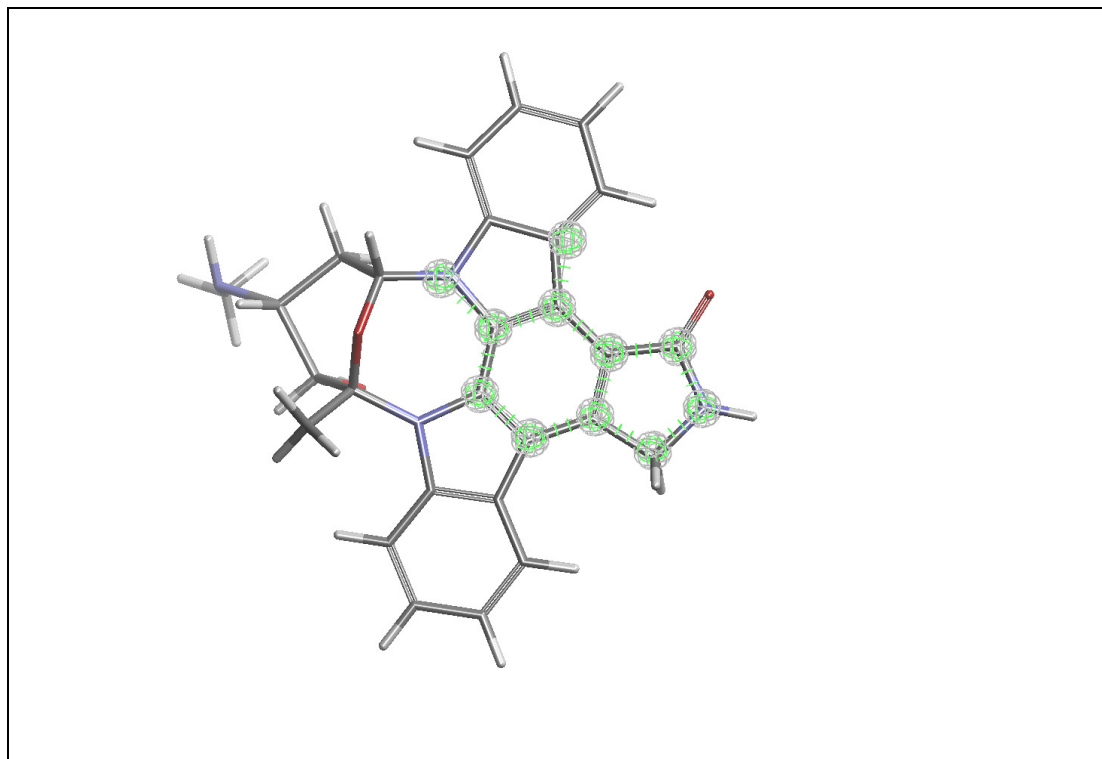




**Figure 10.5** Visualization of D-erythritol 4-phosphate in coot along with environmental distances, the hit totally satisfies the electron density of the unknown ligand in the structure.

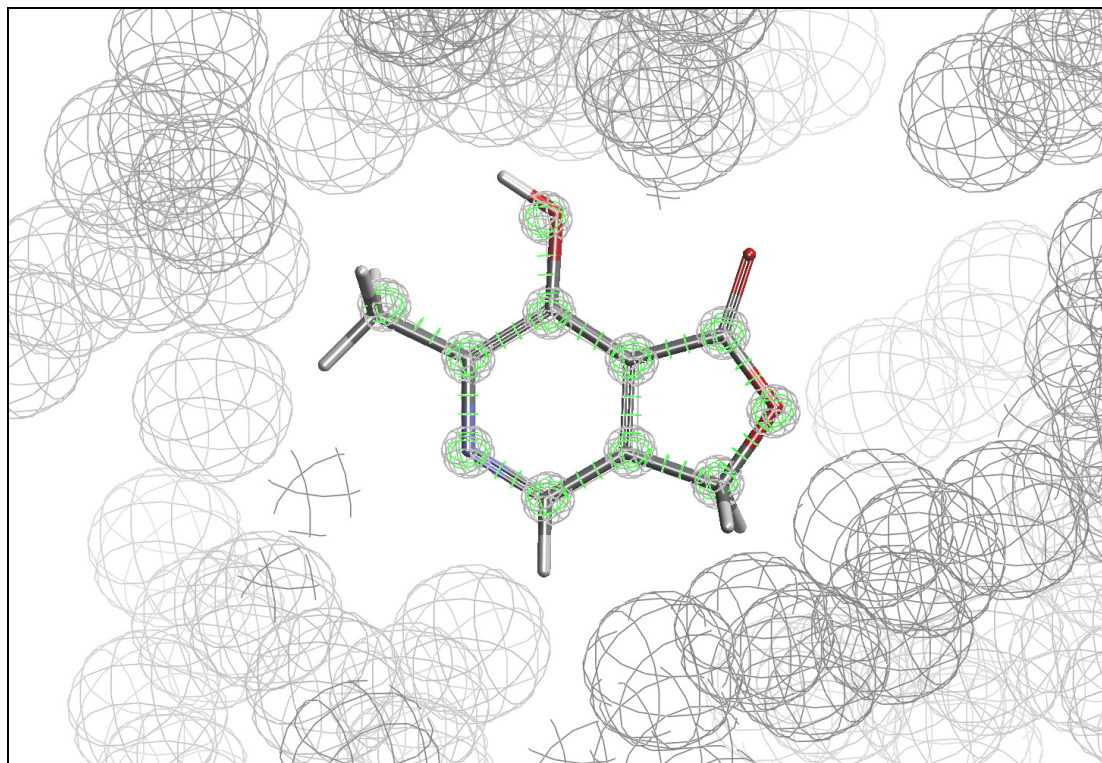
### 10.3.2 3P6K

3P6K is a PLP-dependent aminotransferase (ZP\_03625122.1), found in *Streptococcus suis* 89-1591. The X-ray crystal structure of the enzyme has been determined at a resolution of 2.07Å and submitted under the JCSG project. The enzyme has two chains i.e.; A and B. The unknown ligand with reasonably fair electron density resembled a purine ring. The unknown ligand is present in binding site of the chain B only. The pharmacophore model for the enzyme was created by using the query atom method. In the first pharmacophore only the positions of the individual atoms were specified with bonding connection between them. The hits (59) obtained as a result mainly comprised of big structures like sporines, rifmycin and their derivatives. The hits were clearly and significantly over stepping the query atoms of the pharmacophore and would not be expected to be disordered or fragmented within the crystal structure (Figure 10.6).



**Figure 10.6 Sporines (stick model) as one of the hits clearly violating the limit of the query atoms in the pharmacophore (small grey spheres with green bonding stripes)**

As the hits in the initial pharmacophore were crossing the boundaries in terms of query atoms of the pharmacophore, therefore in the next pharmacophore additional exclusion spheres were introduced around the query atoms (Figure 10.7). This resulted in limited number of hits (4). The hits in this pharmacophore were satisfying the query atoms of the pharmacophore and also were not in steric clash with the surrounding residues e.g.; pyridoxilactone and its derivatives.

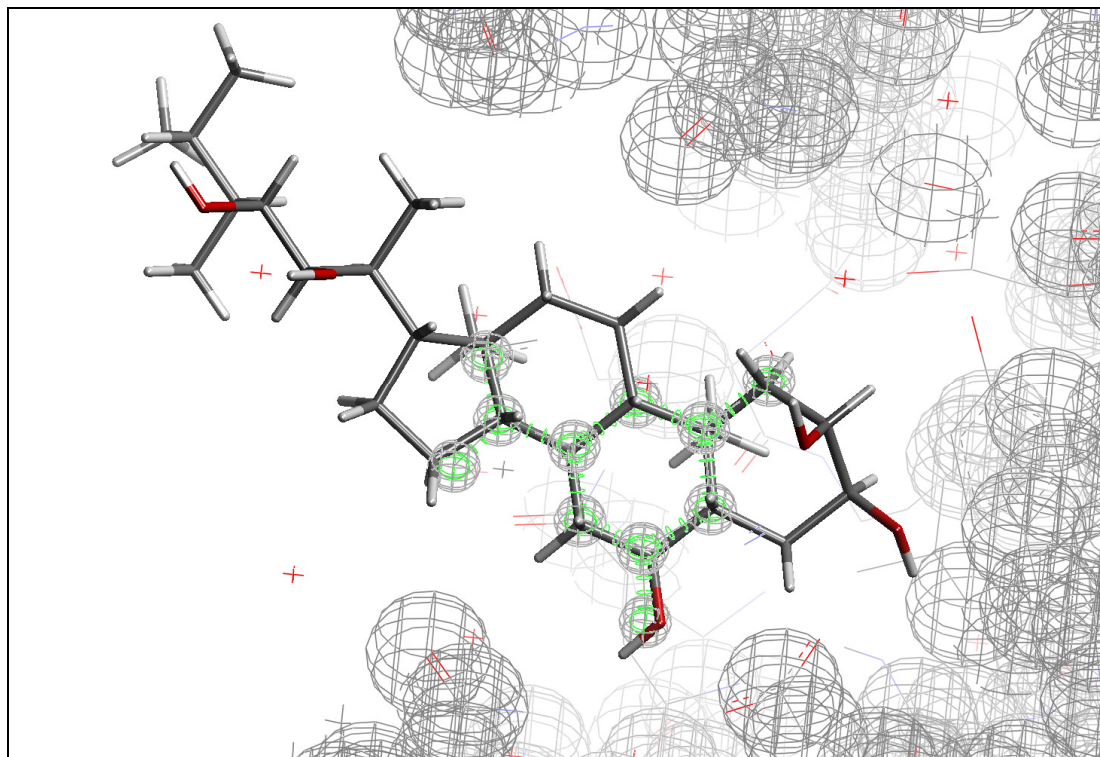


**Figure 10.7** Pyridoxilactone as a hit (stick model) exhibiting compliance with the query atoms of the pharmacophore (small grey spheres with green bonding stripes) the surrounding large grey spheres are the exclusion spheres.

### 10.3.3 2AAM

2AAM is the crystal structure of a putative glycosidase (tm1410) from *thermotoga maritime* diffracted at a resolution of 2.20Å and submitted under the JCSG project. The protein has an unknown ligand with disordered electron density in the binding pocket, present at the same position in all the A, B, C, D, E and F chains of the protein.

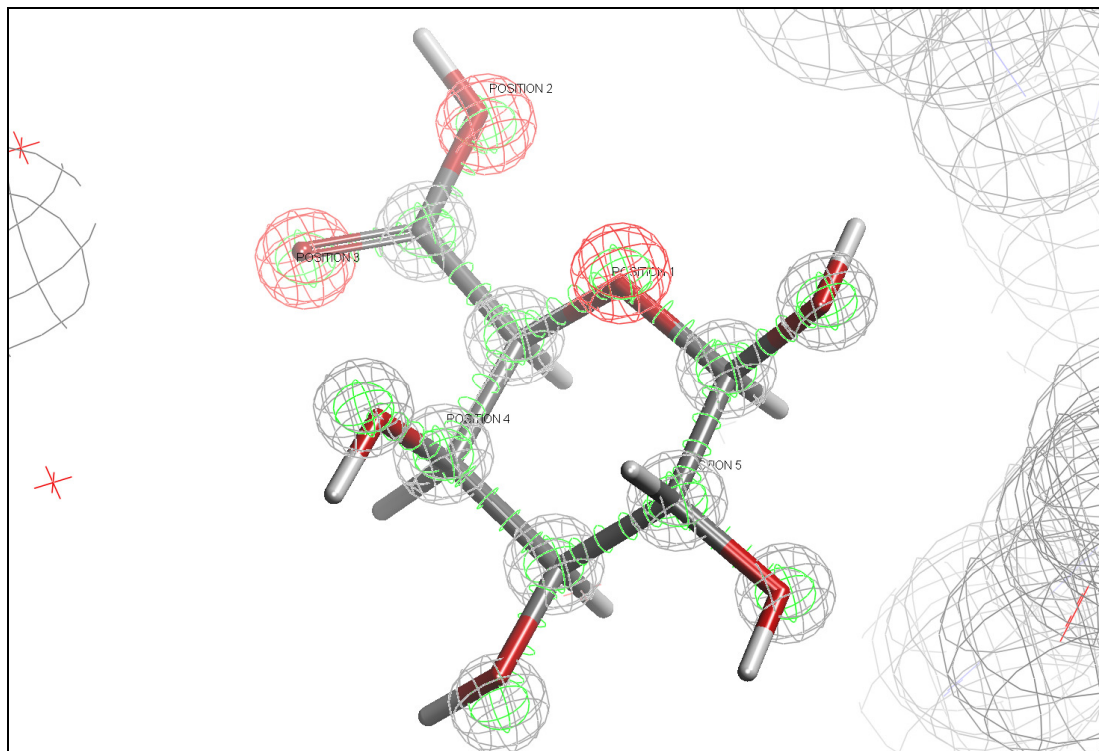
In the initial stages only 11 bonded atoms of the unknown ligand were selected for the generation of pharmacophore model by using the query atom method. Majority of the resultant hits (9) obtained were not in compliance with the model (Figure 10.8).



**Figure 10.8** Initial pharmacophore consisting of 11 bonded atoms (green and grey spheres) based on geometry of the unknown ligand. One of the hit (stick model) shows noncompliance with the pharmacophore model (big grey spheres represent the exclusion spheres)

To further optimize the parameters, in the final pharmacophore model, one of the atoms of the ring and two of the branch were assigned with multiple atoms (C, N, and O) labelled as position 1, 2 and 3 respectively (Figure 10.9). The bonding parameter between these positions was set to either single or double bond by using the Catalyst® pharmacophore optimization parameters. Two additional atoms, bonded to position 4 and 5 were added from the unknown ligand to the pharmacophore model. The number of hits was reduced to nearly half and included beta-D-galacturonic acid. Among other hits beta-D-galacturonic acid was found to be the best hit as it completely satisfied the overall pharmacophore model. When visualized in coot the hit fitted well to the electron density of the unknown ligand.

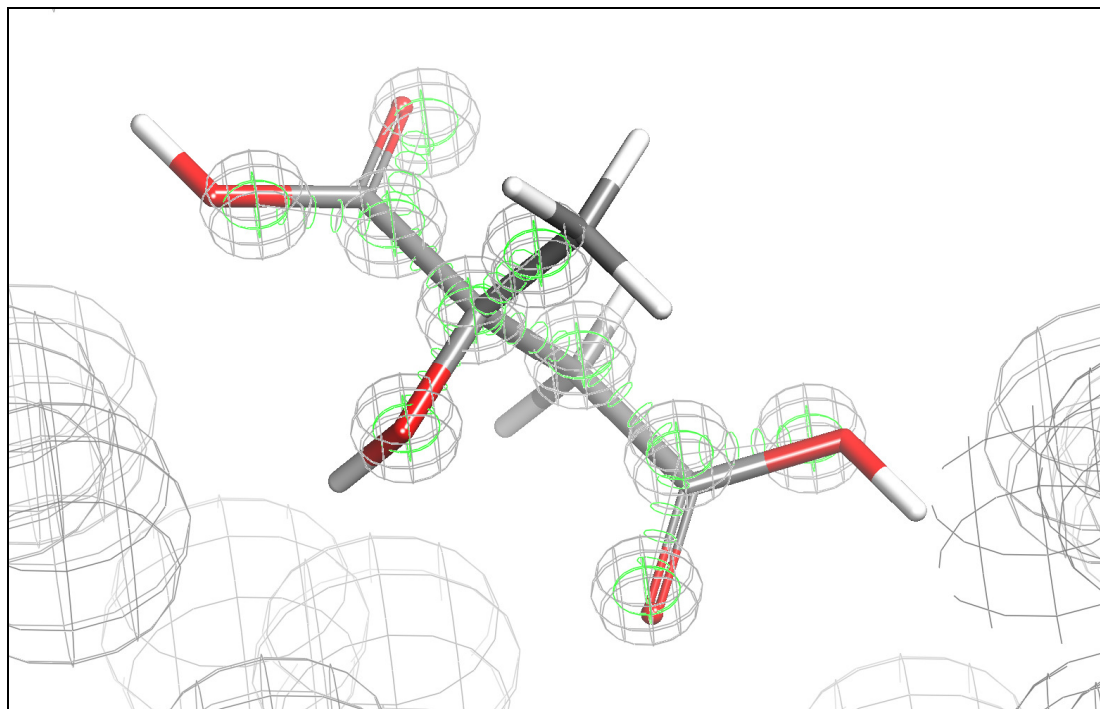




**Figure 10.9** Final pharmacophore with multiply assigned atoms at position 1, 2 and 3, along with additional bonded atoms connected to position 4 and 5. Beta-D-galacturonic acid (stick model) shows complete compliance with the pharmacophore model (certain exclusion spheres and amino acid residues are removed for clear viewing purposes)

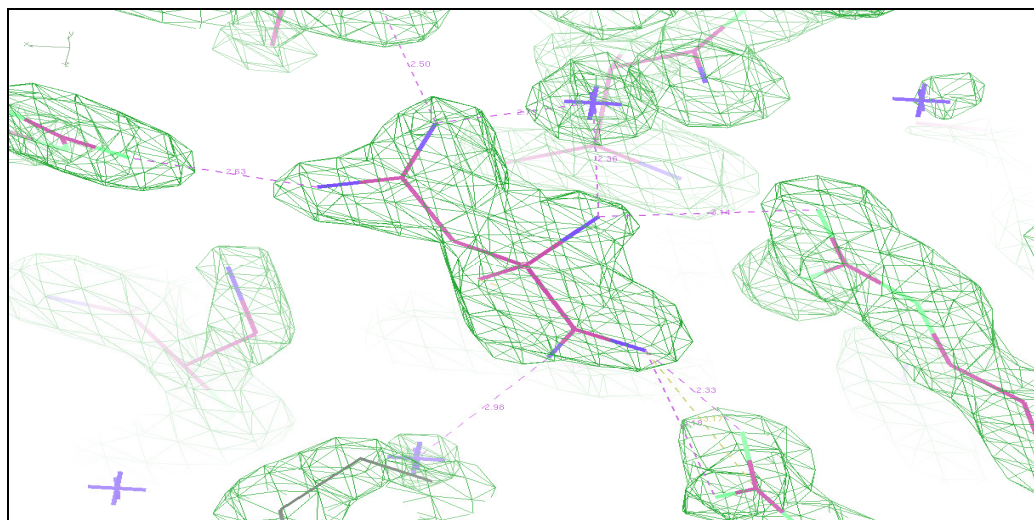
### 10.3.4 2F6R

2F6R is the crystal structure of bifunctional coenzyme A synthase (CoA synthase): (18044849) from *Mus musculus* diffracted at a resolution of 1.70Å, and published under the JCSG project. Based on query atom method pharmacophore model was generated by using the unknown ligand as the template. The model comprised of 10 bonded multiply assigned atoms(C, N and O), with standard valencies. The database search gave convincing hits (35), with majority of them satisfying the maximum pharmacophore parameters. After visual inspection L-citramalic acid came out as a valid hit (Figure 10.10) which was absent in the crystallization conditions. This further increases its potential as a true ligand of the enzyme. Citric acid and citrate were also among the hits, which can be the putative ligands.



**Figure 10.10** L-citramalic acid (stick model) among the hits, fully satisfies the introduced constraints in the pharmacophore model. The small green and grey spheres are the multiply assigned pharmacophore atoms, the big grey spheres around the ligand are exclusion volume spheres.

By visualizing the hits in coot it appeared that the methyl group of citramalic acid fitted well into the partially disordered electron density region (Figure 10.11). Keeping in view the disordered electron density no further constraints were applied as it can cause to loose potential hits.



**Figure 10.11** Graphical representation of L-citramalic acid in coot along with environmental distances, the hit is in agreement with the electron density of the unknown ligand in the structure (the hydrogen atoms have been removed for clarity viewing)

### 10.3.5 1TVF

1TVF is the crystal structure of penicillin-binding protein 4 (PBP4) from *Staphylococcus aureus*. The unknown ligand has been defined in terms of individual non-bonded atoms with good electron density around each atom of the unknown ligand. The pharmacophore was generated by using the query atom method. The individual atoms of the unknown ligand were assigned as multiple atoms (C, N and O) in the model. The majority of the hits (10) obtained through database search included citrate and its derivatives. Interestingly in the crystallization conditions of the protein citrate is used as a buffer, which implies that the unknown ligand most likely to be a citrate molecule, being crystallized along with the protein during crystal formation. Figure 10.12-14 show the schematic progress of pharmacophore generation and the outcome from the database search. In figure 10.12 the individual atoms of the unknown ligand are labelled as “unknown ligand atoms”. In figure 10.13 these atoms are specified as the query atoms of the pharmacophore in the binding site of the protein. In figure 10.14 citric acid satisfies all the constraints of the pharmacophore and fits well in the binding site and is most likely to be a true ligand.

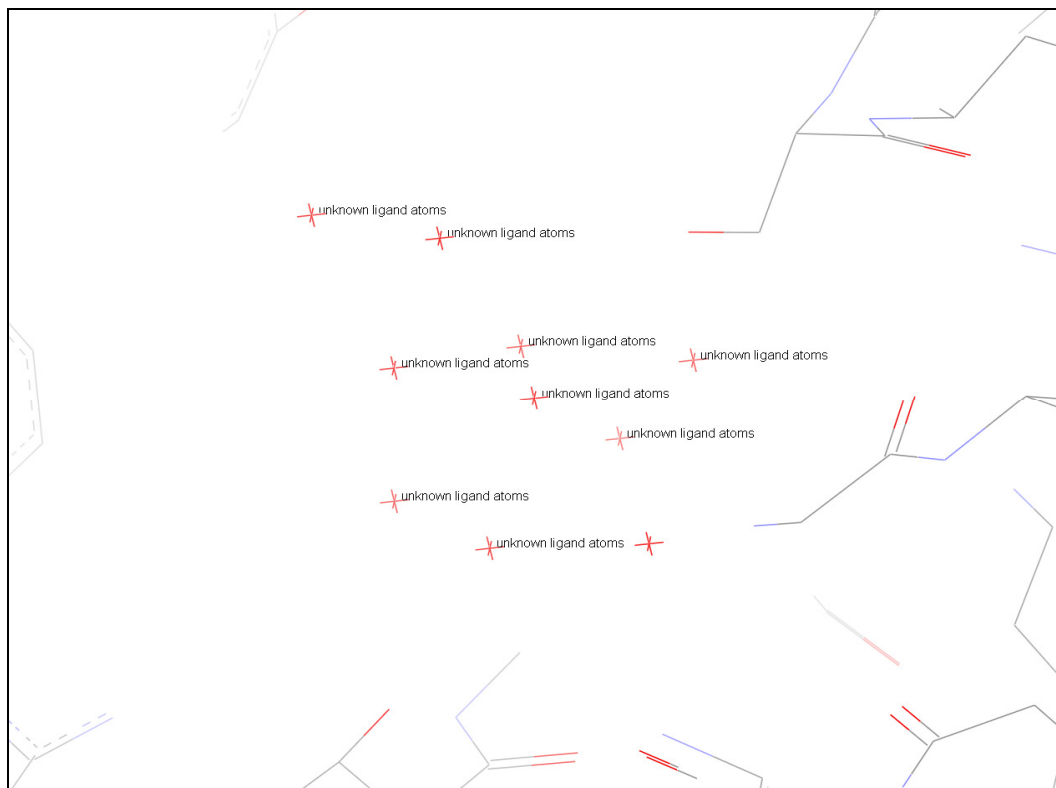
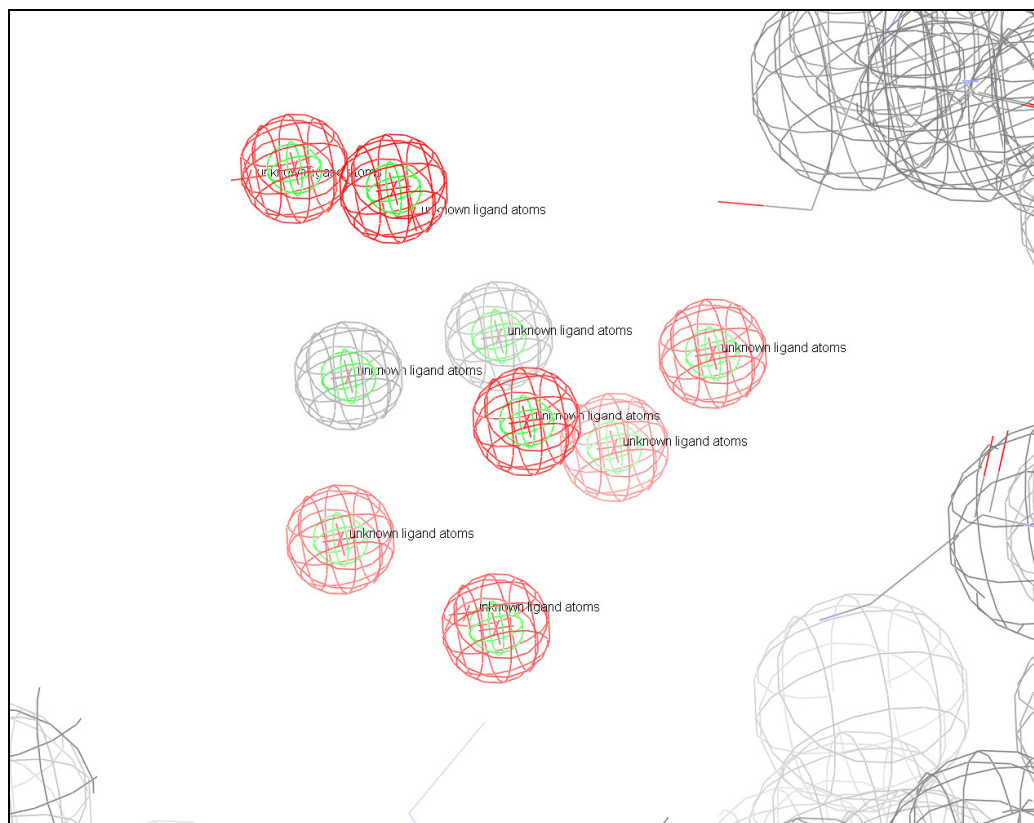
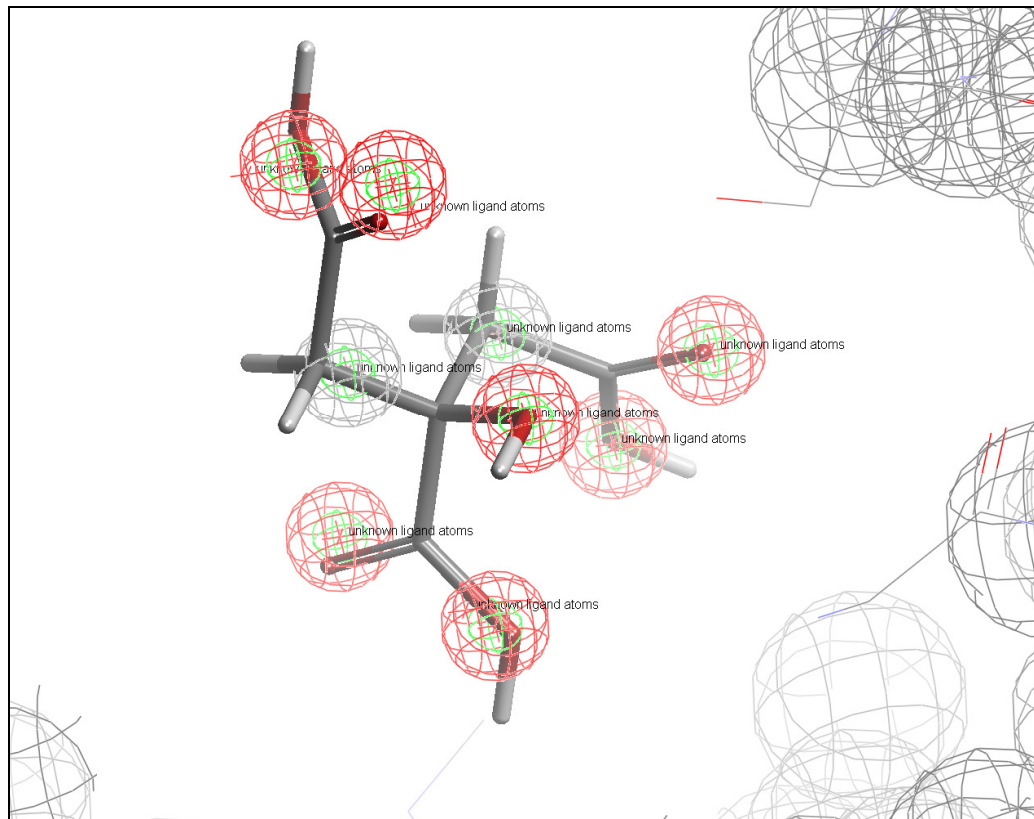


Figure 10.12 Selection and labelling of unknown ligand atoms for pharmacophore model



**Figure 10.13** Query atoms of the pharmacophore along with exclusion spheres in binding site

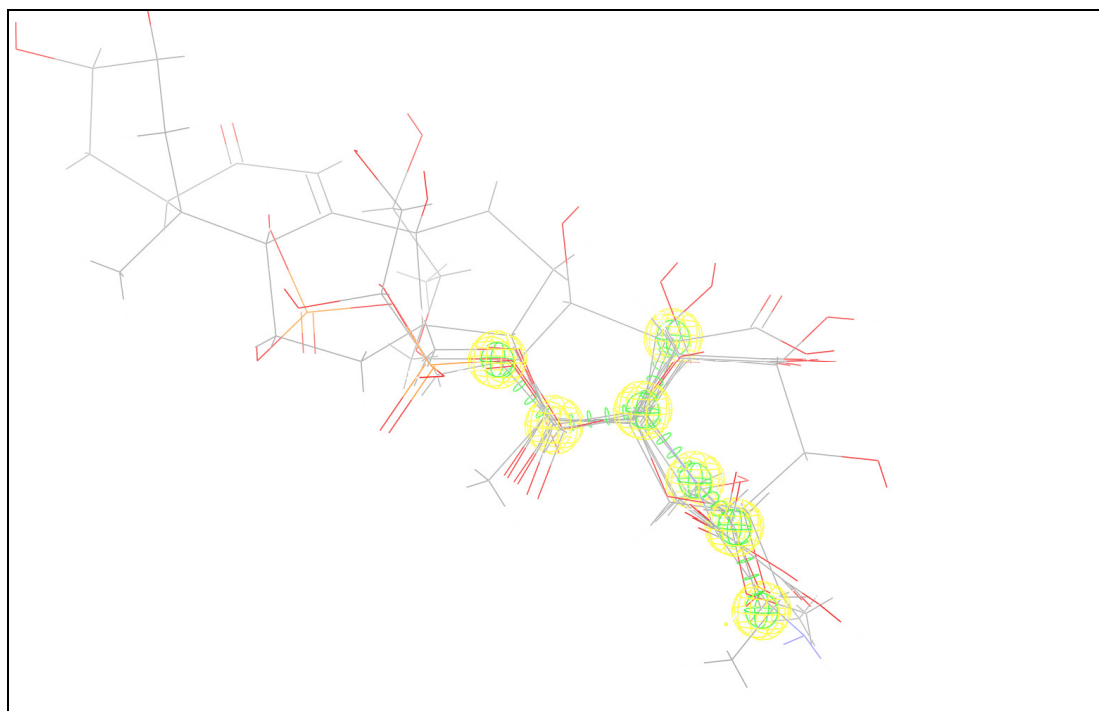


**Figure 10.14** Citric acid as a hit satisfying all the constraints of pharmacophore model in the binding pocket of the protein



### 10.3.6 2PY6

2PY6 is the crystal structure of methyltransferase FkbM (YP\_546752.1) from *Methylobacillus flagellatus* KT diffracted at a resolution of 2.20Å. The structure is deposited under the JCSG project. The unknown ligand is situated well inside the binding pocket. The pharmacophore model for the protein was generated by using query atom method. Most of the hits obtained through pharmacophore searching of the database were penetrating beyond the specified query atoms of the model and thus violating the provided constraints (Figure 10.15). A possible reason for this problem can be the water molecules in the protein structure, around which the exclusion spheres are not introduced in the pharmacophore model and thus the software logic may consider those spaces as free, thus resulting in selection of longer and bigger ligand structures or this can be due to a flaw in the catalyst software. Apart from violating the pharmacophore model, some of the hits satisfied the model, which included some sugars and amino acids like L-homoserine, erythritol, L-erythrulose and L-arabinose.



**Figure 10.15** Majority of the hits (line models) clearly violating the atomic constraints of the pharmacophore model (highlighted small yellow spheres and green bonding stripes, amino acid residues and exclusion spheres are removed for clarity purposes)

### 10.3.7 3EZU

3EZU is the crystal structure of multidomain protein of unknown function with GGDEF-domain (NP\_951600.1) from *geobacter sulfurreducens* diffracted at a resolution of 1.95Å. The structure is deposited under the JCSG project. The unknown ligand had good electron density and situated well within the binding site. The pharmacophore model was generated by using the query atom method. The bonded query atoms along with the exclusion spheres were then subjected to the database search. The search gave hits, among which many satisfied the pharmacophore completely and some violated the pharmacophore. Few hits like 2-ethylhexan-1-ol, triethanolamine and diethylaminoethanol (Figure 10.16) fitted the electron density and are most likely to be potential ligands of the protein.

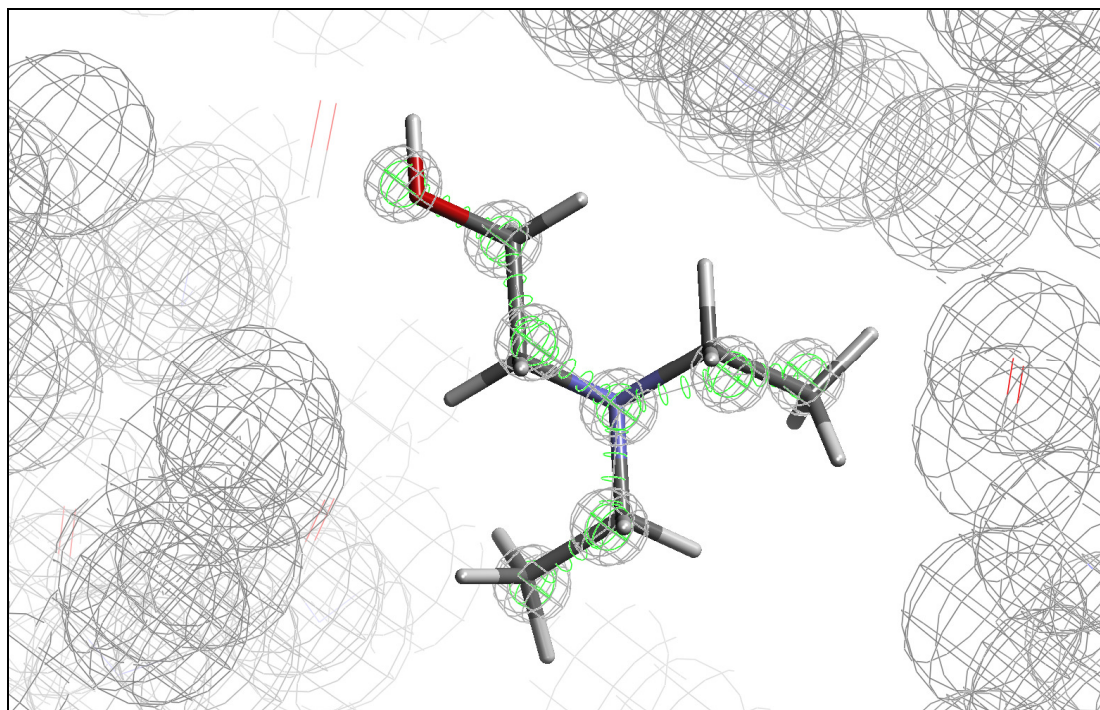
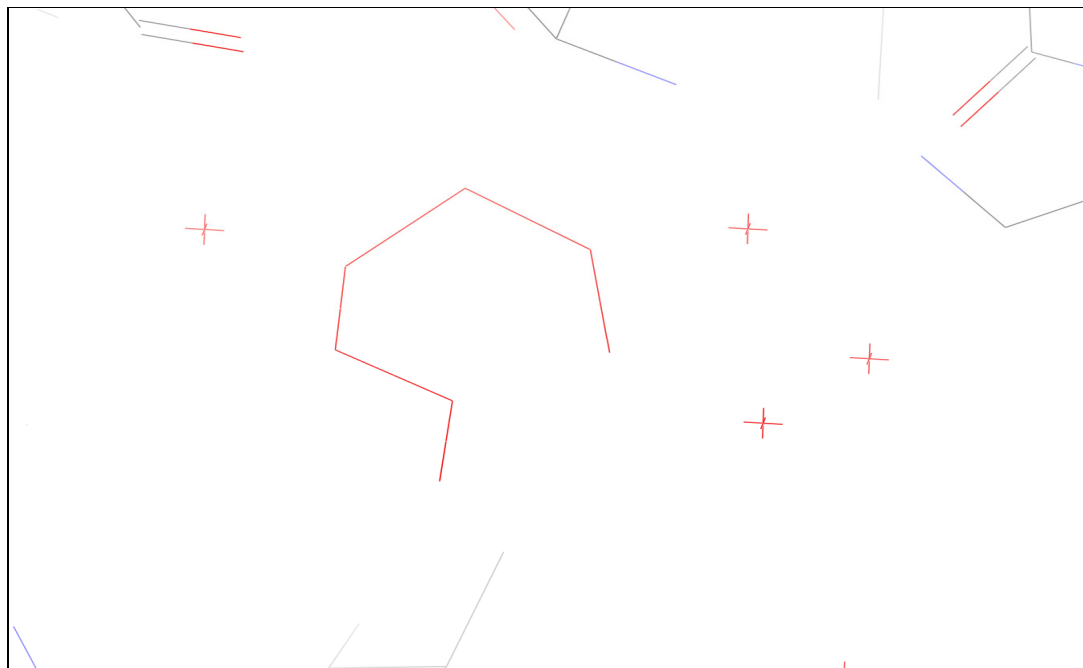


Figure 10.16 Diethyl amino ethanol fits the pharmacophore model and the electron density map of the unknown ligand in the binding site of the protein.

### 10.3.8 1VK8

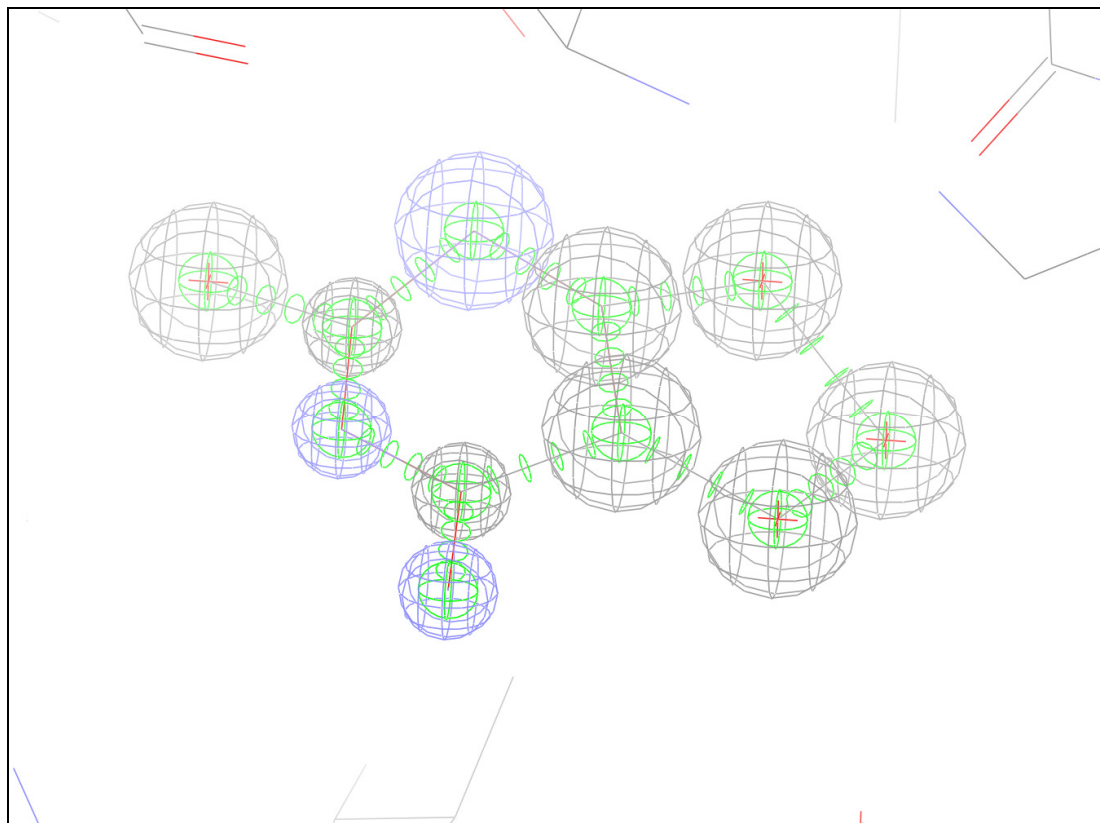
1VK8 is the crystal structure of a putative thiamine biosynthesis/salvage protein (tm0486) from hyperthermophilic anaerobe *Thermotoga maritima* diffracted at a resolution of 1.80Å, published in journal of molecular biology {172}. Protein belongs to a family of thiamin-binding proteins which are involved in responding to the cell oxidative conditions. The protein has a ferredoxin-like fold and is a

homo-tetramer. The unknown ligand was identified in the pocket of each monomer, with well defined electron density. Some atoms of the unknown ligand are bonded to each other, while some are identified as individual entities, lying close to the bonded unknown atoms (Figure 10.17).



**Figure 10.17 Bonded atoms (line model in red) of the unknown ligand and non-bonded atoms (red stars) of the unknown ligand in the binding site of the protein.**

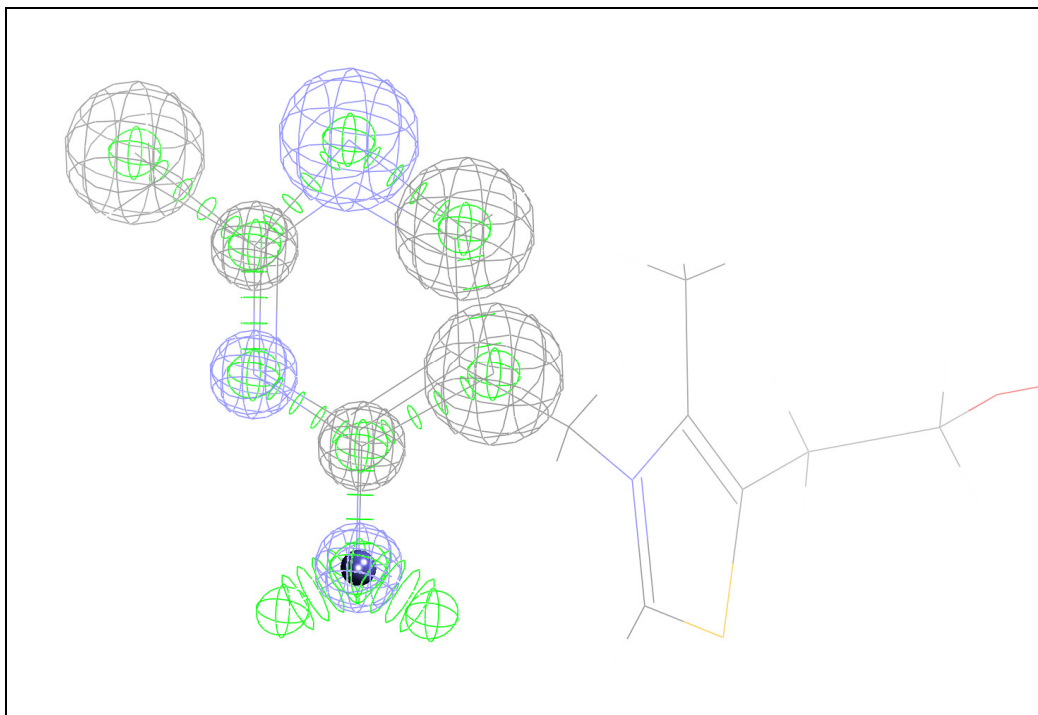
A number of pharmacophore models were generated for the binding site of the protein by using the query atom method. In the first pharmacophore model, all the bonded and non bonded atoms of the unknown ligand were used as a template for the query atoms. The model was optimized by specifying the bonds between the non-bonded atoms and bonds between non-bonded and bonded atoms by using the Catalyst software. As the protein is a thiamine binding protein, therefore the atom specification (C,N and O) and bond type (single, double) were specified in the pharmacophore model in relation to thiamine (Figure 10.18). The hits obtained as a result did not have thiamine among them.



**Figure 10.18 Pharmacophore model comprising of query atoms with respect to the position of unknown ligand atoms (both bonded and non-bonded) in the binding site of the protein (exclusion spheres are removed for clear viewing)**

To further improve the pharmacophore in terms of pulling thiamine as a hit, in the new model the query atoms around the non-bonded atoms of the unknown ligand were removed. A total of 166 hits were obtained through this model which included thiamine and its corresponding aldehyde, carboxylic and phosphate forms. This showed that the incorporation of non-bonded atoms of the unknown ligand in the model is not helpful for sole selection of true ligand of the protein.

In order to reduce the number of hits without losing the potential ligands, the hydrogen count on the nitrogen query atom was set to two in the final model (Figure 10.19). This resulted in only 35 hits which included the thiamine and its aldehyde, carboxylic and phosphate forms, exhibiting that specifying the hydrogen count on the nitrogen atom serves to reduce the hits to a reasonable number without losing the potential ligands.

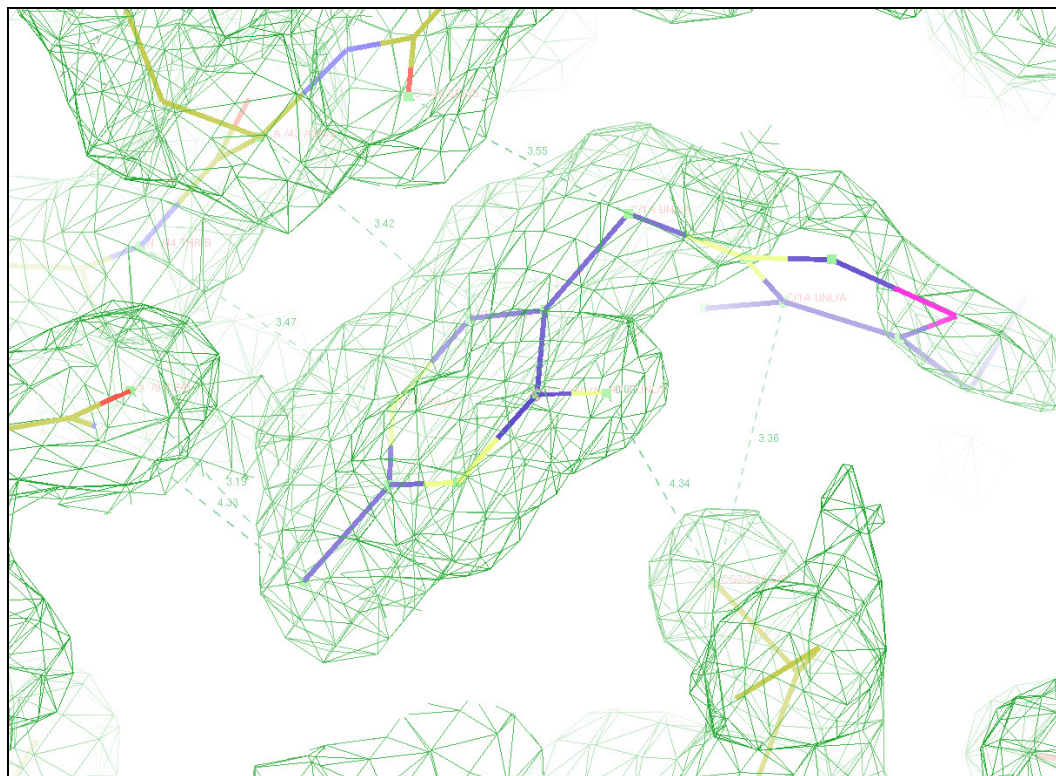


**Figure 10.19** Final pharmacophore model with non-bonded atoms removed and the hydrogen count on nitrogen (solid blue sphere) fixed to two, the thiamine (line model) fits well in the pharmacophore without clash with the surrounding exclusion spheres in the binding site (exclusion spheres are hidden for clarity purposes)

It will be interesting to mention here that in pharmacophore modeling, exclusion spheres are introduced by selecting area of the binding site with certain radius (Å). The selection excludes the coordinating water molecules in the protein structure. This provides more 3D space and flexibility in search for selecting comparatively bigger ligands, as in case of 1VK8 (thiamine binding protein), where during the search the availability of bigger 3D space made it possible to identify and accommodate comparatively bigger ligand, thus thiamine and its derivatives were selected among other hits.

By visualizing the hits in coot it appeared that the aminopyrimidine ring of thiamine fitted well into the electron density while the thiazole ring fitted in the partially disordered electron density region (Figure 10.20). Keeping in view the disordered electron density no further constraints were applied as it could cause to loose the potential hits.

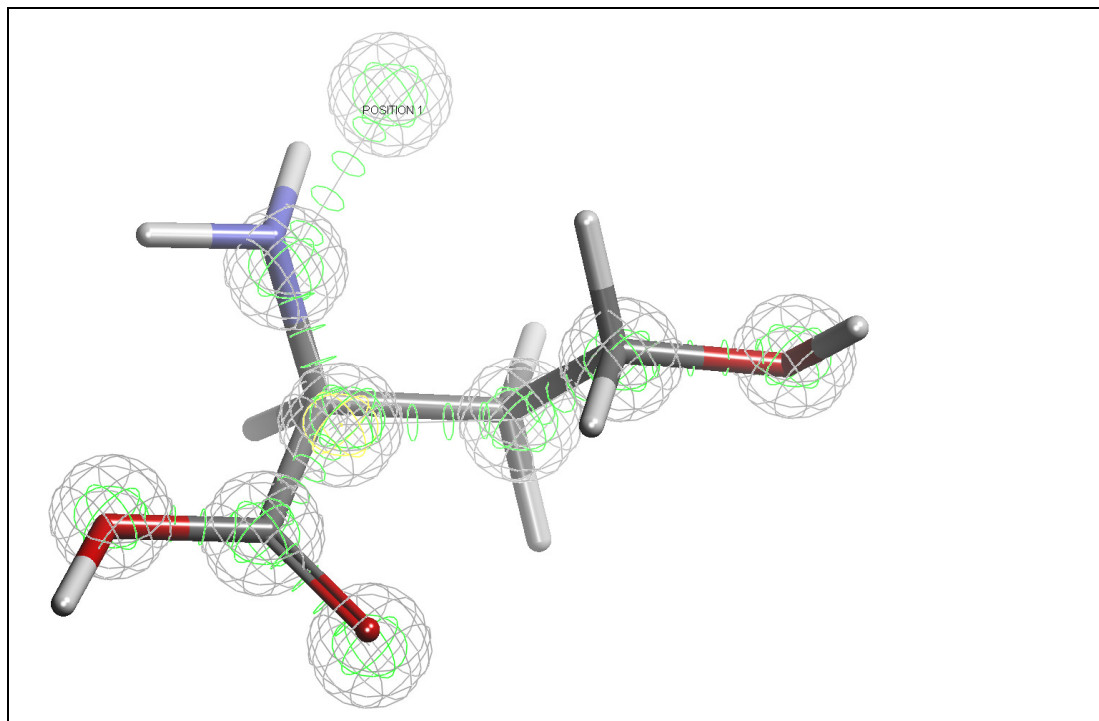




**Figure 10.20** Graphical representation of thiamine in coot along with environmental distances, the hit is mostly in agreement with the electron density of the unknown ligand in the structure (the hydrogen atoms have been removed for clarity viewing).

### 10.3.9 2FTR

2FTR is the crystal structure of an ethyl tert-butyl ether d (ethd) family protein (bh0200) from *bacillus halodurans* c-125, diffracted at a high resolution of 1.40Å. The structure is submitted under the JCSG project. The unknown ligand is located well inside the binding pocket of the protein with reasonably good electron density. The pharmacophore model for the protein binding site along with the unknown ligand as a template was generated by using the query atom method. The initial pharmacophore model gave hits, which were not aligning properly with the model. Therefore in the final model one of the query atoms at position one was removed (Figure 10.21). The modification in the model resulted in partial freedom and thus selection of suitable ligands in response. Some the convincing hits included compounds like homo-serine, L-leucine, L-homoserine, creatine, L-aspartic acid, L-asparagine, oxaloacetic acid, acetylpyruvic acid and malic acid. These hits were not pulled out by the previous pharmacophore and can be the potential ligands of the protein.

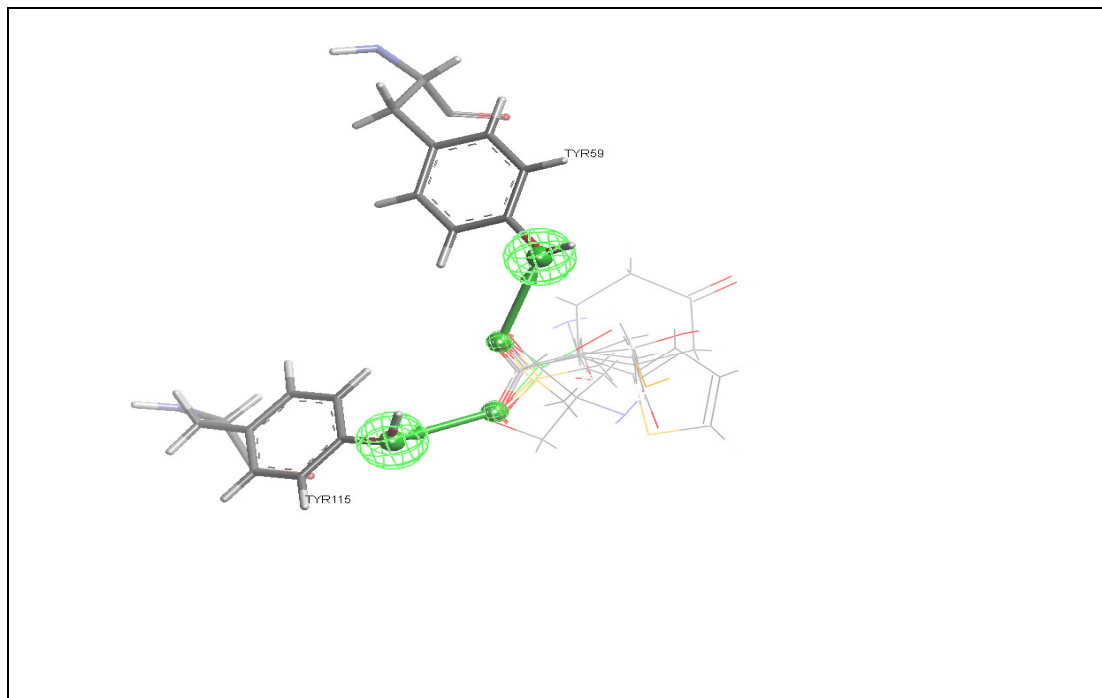


**Figure 10.21** Pharmacophore model based on unknown ligand in the binding site of the protein. Position 1 was removed in the final model, allowing more flexibility in selection, among the hits homoserine (stick model) best fitted the model (amino acid residues and exclusion spheres are hidden for clarity)

### 10.3.10 3NNR

3NNR crystal structure has x-ray data to a resolution of 2.49Å, which belongs to TetR-family of protein a transcriptional regulator (Maqu\_3571) from *Marinobacter aquaeolei* VT8. The X-ray structure is published under the JCSG project. The assigned unknown ligand is kind of a long chain with good electron density and situated well within the binding site of the protein. The resolution is not particularly high and as such represents a challenging case.

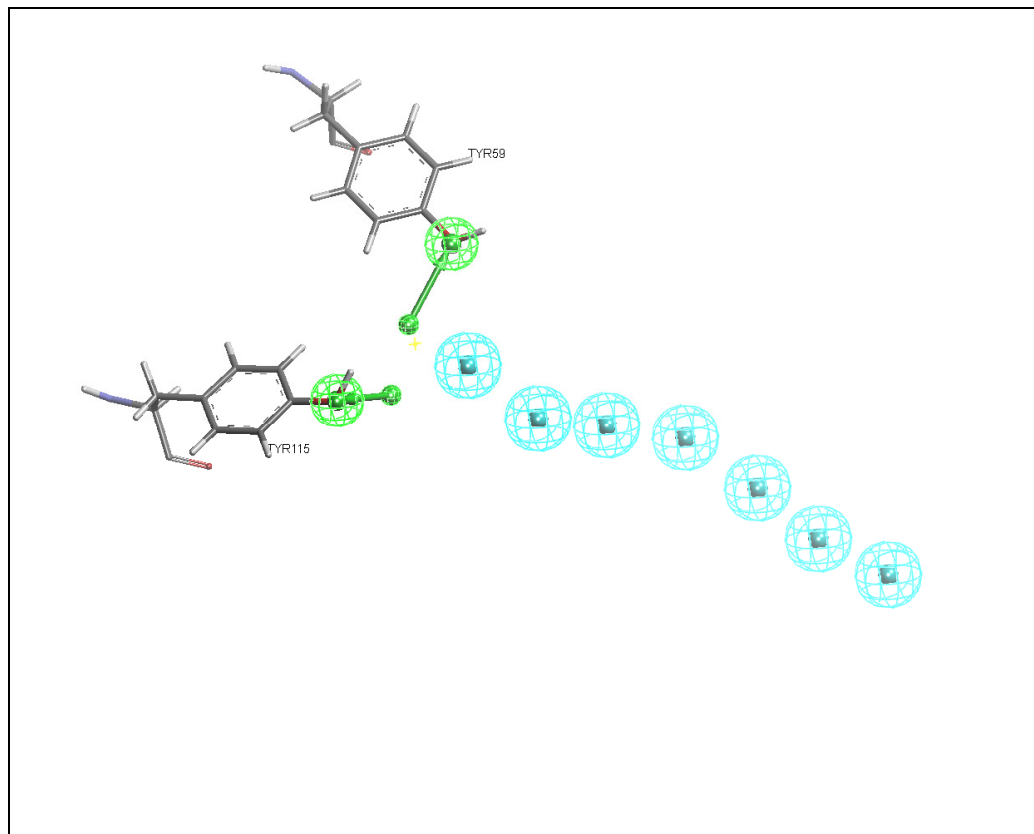
A number of pharmacophores were generated by using the H-bond vector method. In the first model two H-bond acceptor vectors were introduced from the hydroxyl group of Tyr59 and Tyr115, majority of the hits (210) obtained through this model were mainly oriented and confined in the surrounding area of Tyr59 and Tyr115 (Figure 10.22). The hits included various amino acids, short chain sugars. However a few hits like lauric acid were able to partially satisfy the electron density of the unknown ligand in the form of a long chain fatty acid.



**Figure 10.22 Pharmacophore model with 2 H-bond acceptors from Tyr59 and Tyr115 in the active site, majority of the hits shown as line models are confined in the surroundings of Tyr59 and Try115.**

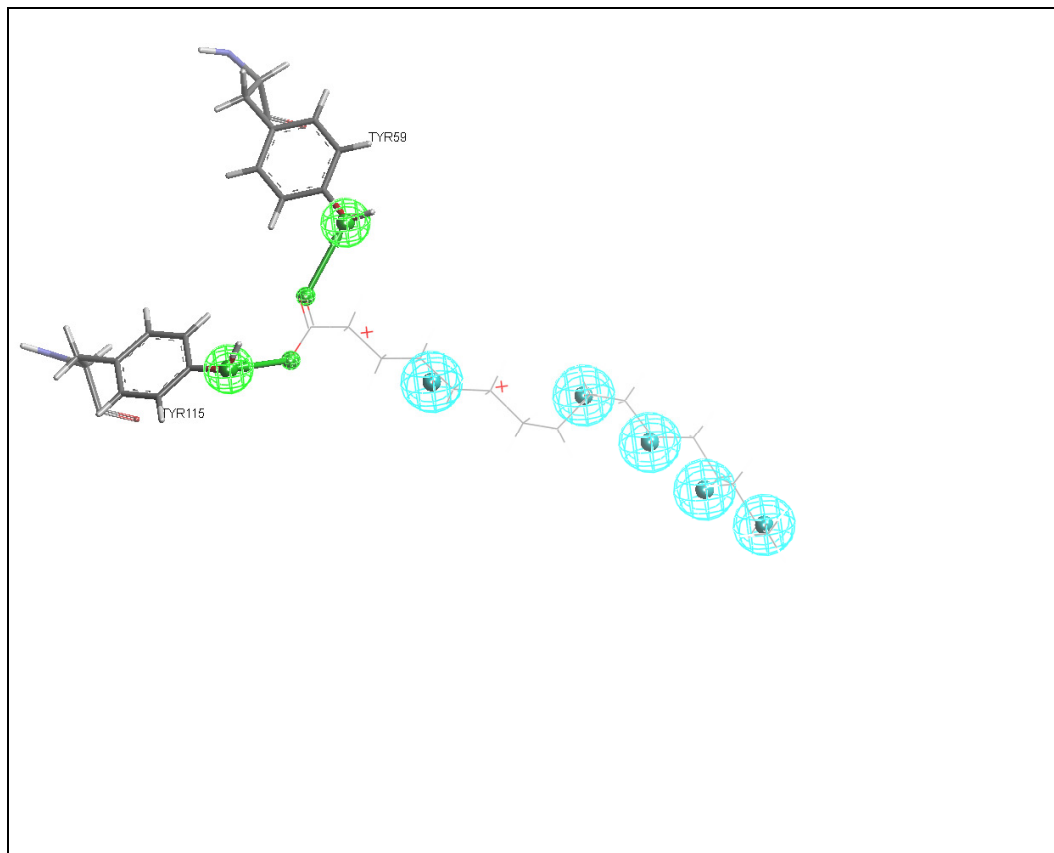
Keeping in view the presence of long chain electron density in the binding pocket the 2<sup>nd</sup> pharmacophore model, in addition to the two H-bond acceptor interactions, the model was further modified by specifying the individual atoms of the unknown ligand as 7 hydrophobic points (Figure 10.23). The hits obtained (300) as a result of this model were quite encouraging as maximum hits were long chain fatty acids, with their carboxylic groups pointing towards the H-bond acceptor vectors of Tyr115 and Tyr59 and the corresponding hydrophobic chains were positioned exactly towards the hydrophobic points created in the model.





**Figure 10.23** Introduction of hydrophobic points (blue spheres) at the designated positions of the individual atoms of the unknown ligand specified in X-ray structure.

To further explore the binding site, the numbers of hydrophobic points were reduced to 5 in the pharmacophore model (Figure 10.24). The resultant hits obtained from the 3<sup>rd</sup> pharmacophore model were only 8. The hits included palmitic acid, linoleic acid and their derivatives. This indicated that the specific number and position of hydrophobic points in the model is crucial for the number and type of hits.

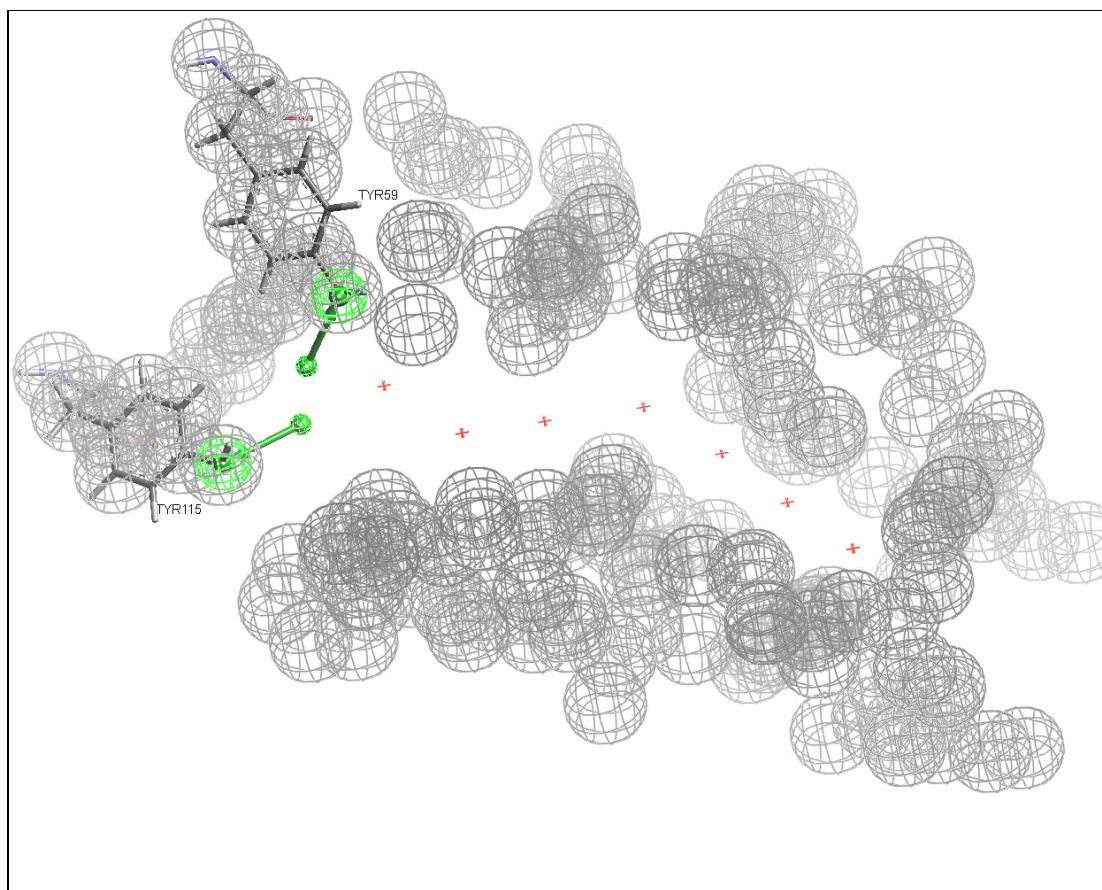


**Figure 10.24** 2 H-bond acceptor (green arrows) and 5 hydrophobic points (cyan spheres) in the pharmacophore model along with palmitic acid as a hit shown in line model in the binding site of the protein.

In the 4<sup>th</sup> pharmacophore model one of the hydrophobic points was removed from the very end and instead a hydrophobic point was introduced between the 1<sup>st</sup> and 2<sup>nd</sup> hydrophobic points. The number of hits (17) was doubled as a result and included some other fatty acids like mauristic acid and its derivatives while palmitic acid was not included in the hits, showing that the removed hydrophobic point at the very end is essential for the selection of palmitic acid and its derivatives.

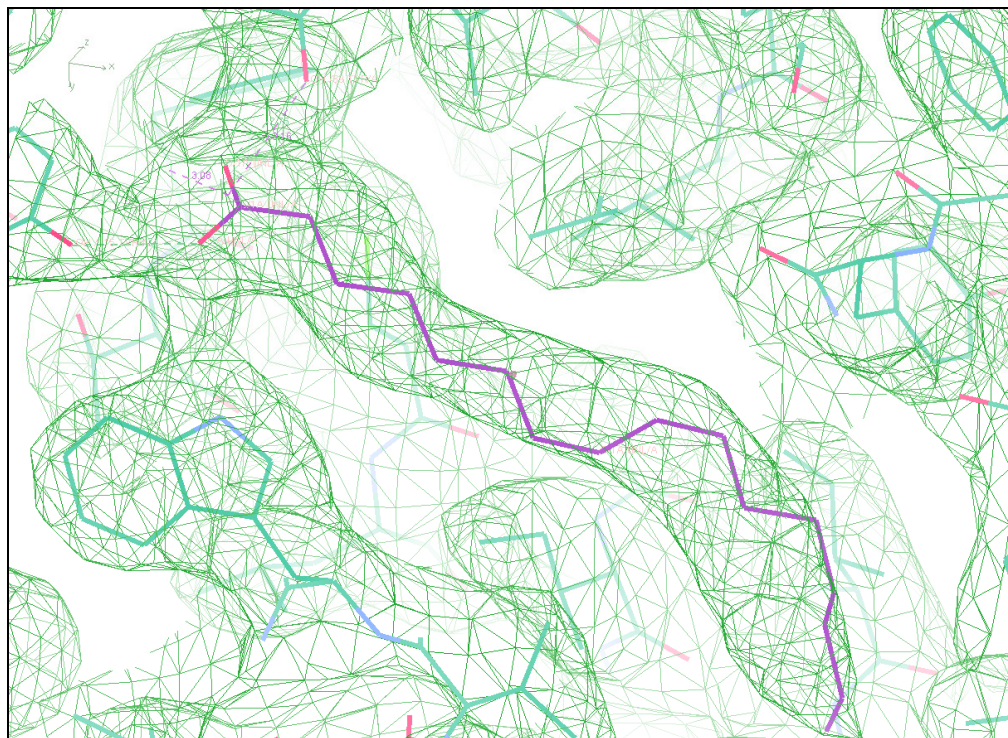
To test whether the presence of hydrophobic points is essential in the pharmacophore model for selection of long chain fatty acids as hits, all the hydrophobic points were removed and the individual atoms corresponding to the unknown ligands were selected for generation of exclusion spheres around them (Figure 10.25). The hits (210) obtained included many long chain fatty acids like capric acid, decanoic acid and dedecanoic acid and their derivatives, showing

that the presence of hydrophobic points is not crucial for selection of long chain fatty acids.



**Figure 10.25** Pharmacophore model generated with 2 H-bond acceptor interactions along with exclusion spheres specified around the designated atoms of the unknown ligand.

The visualization of hits in coot indicated that mostly the hits containing long chain fatty acid satisfy the electron density. By close observance it appeared that palmitic acid best fits the electron density (Figure 10.26). Therefore, the conclusion can be drawn that the most likely ligand is a fatty acid in the form of palmitic acid.



**Figure 10.26** Graphical representation of palmitic acid in coot along with environmental distances, the hit is in reasonable agreement with the electron density of the unknown ligand in the structure (the hydrogen atoms have been removed for clarity viewing)

## 10.4 Conclusions

After evaluating various PDB structures through various stages of visualization and modelling search, the following opinions are made

1. The availability of good electron density for the unknown ligand is a major factor in selection for pharmacophore searching and getting sensible hits. X-ray resolution of 2.5Å or better was necessary for the placement of atomic positions, whilst hydrogen bonding vectors could be used in lower resolution and other restraints would be necessary.
2. The structures with disordered and/or poor electron density of the unknown ligand are unlikely for pharmacophore modelling.
3. Labelling of unknown ligands is varied and random, as in most cases a single atom has been labelled as an unknown ligand, which most likely can be an ion or a water molecule.

4. In case of 1VKM , D-erythritol 4 phosphate which is obtained as a potential hit and best fits the electron density, can be used experimentally to evaluate ligand binding.
5. The pharmacophore method works successfully to find the suitable ligand even if the bonding electron density is absent between the atoms, as observed in case of 1TVF.
6. The pharmacophore searching works efficiently in finding logical ligand for the target protein i.e. palmitic acid in case of 3NNR.
7. PDB structures can be used to identify and/or suggest potential ligands for proteins in some cases but not on the whole because of limiting factors like electron density, hydrophobic pocket, ligand at the periphery and very small ligand.
8. To obtain sensible hits certain parameters have to be optimized during the course of searching, leading to freedom in selection of hits. DSV and Catalyst have the option to parameterize certain features of the pharmacophore i.e. type of bond (double or single), Hydrogen count on atoms (1, 2 or free) and assigning multiple atomicity (C, N, S, O) to the query atoms.

On the basis of outcome obtained from above examples, it can be concluded that pharmacophore searching works in a significant way to find a suitable ligand both for a known and unknown protein. The hits obtained from the search could indicate the crucial H-bond interactions which are necessary for the ligand binding event. The method can be readily used for searching specific classes of compounds to be potential ligand for certain proteins. In order to find out the protein ligand interaction in detail the hits obtained from pharmacophore searching can be subjected to ligplot. Ligplot can give information with regard to the hydrogen bond interactions, hydrophobic interactions between the ligand and the amino acid residues of the protein active site. For instance in case of 1VKM, ligplot gives detail interaction map between D-erythritol 4-phosphate and the binding site of the protein (Figure 10.27).

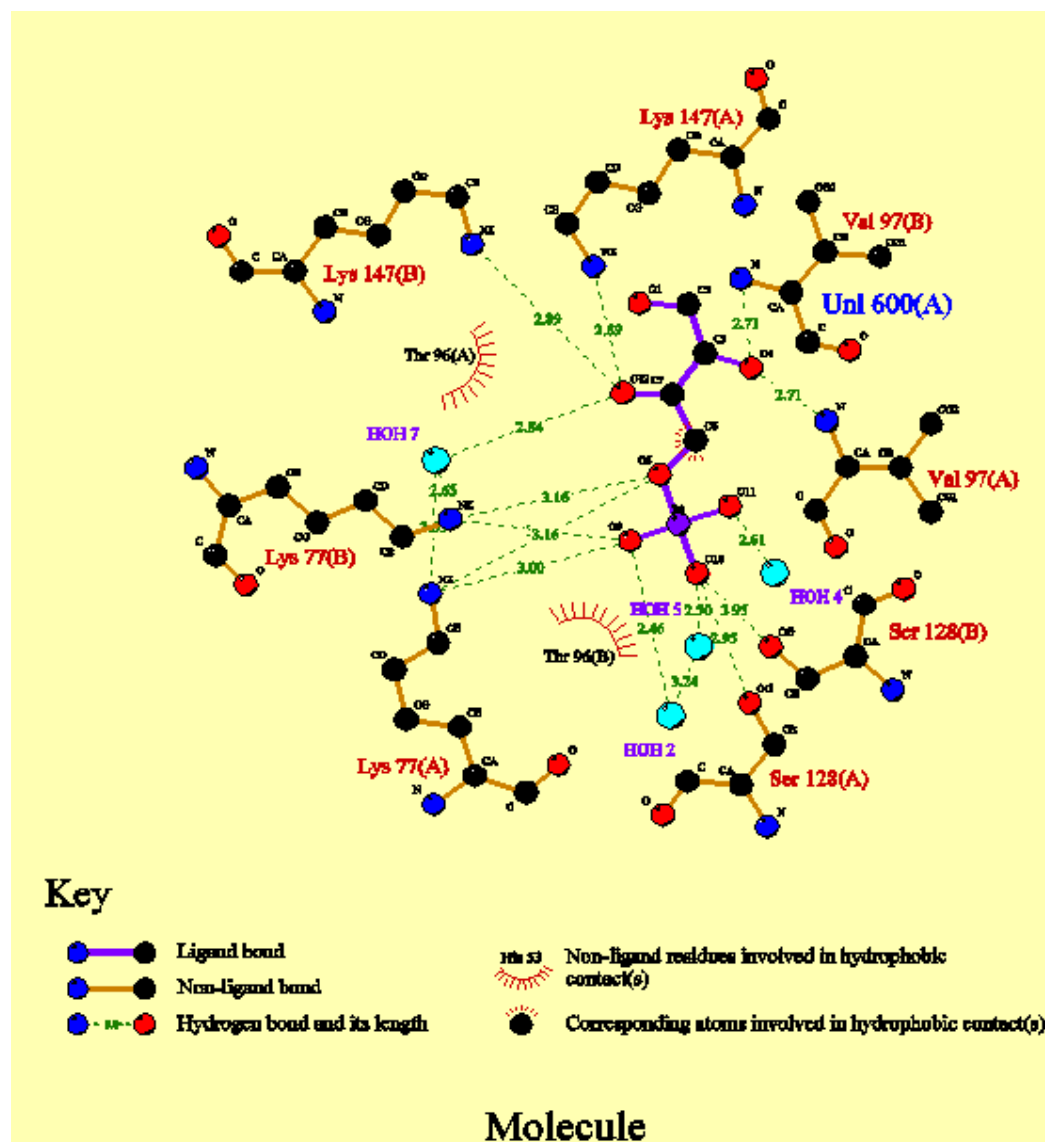


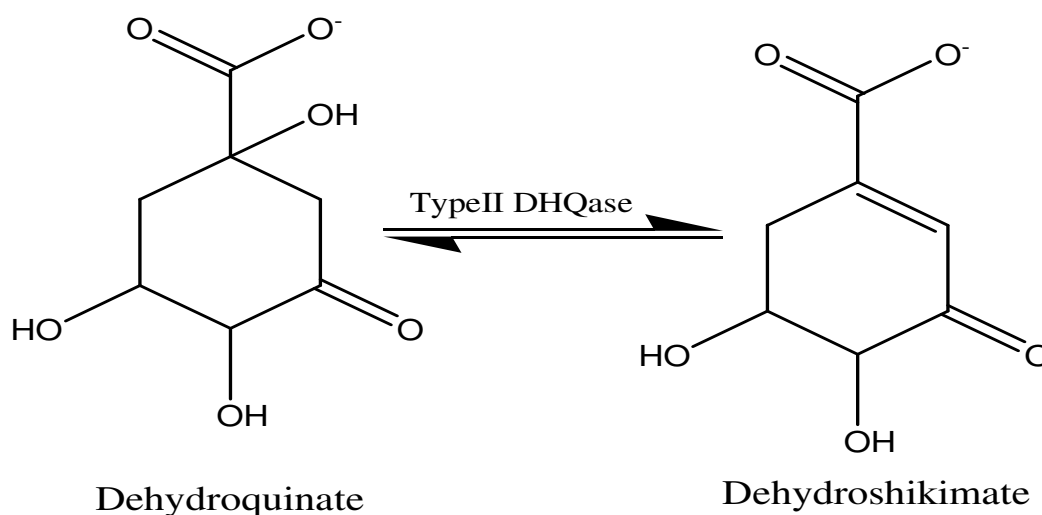
Figure 10.27 A typical ligplot representing the protein-ligand interactions between D-erythritol 4-phosphate and amino acid residues in the active site of 1VKM.

The following Publications have recently been published from this work

1. Rapid assembly of potent type II dehydroquinase inhibitors via "Click" chemistry, Medchemcomm,P 271-275, D.O.I:10.1039/c0md00097c, Oct, 2010 {173}
2. Synthesis and evaluation of potent ene-yne inhibitors of type II dehydroquinases as tuberculosis drug leads, Chemmedchem,P 262-5, D.O.I:10.1002/cmdc.201000399, Feb, 2011 {174}

## 11. Type II DHQases

The enzyme dehydroquinase (3-dehydroquinase dehydrates; DHQase) catalyses the reversible dehydration of dehydroquinic acid (Figure 12.1) to form dehydroshikimic acid {175}. The reaction is part of the two metabolic pathways, the biosynthetic shikimate pathway and the catabolic quinate pathway {176, 177}. The shikimate pathway is an essential biosynthetic pathway involved in the biosynthesis of aromatic amino acids in many microorganisms but is absent in humans {142}. The dehydration of dehydroquinase is catalyzed by two distinct classes of dehydroquinase (type I and II) by different mechanisms {178} . Both types have marked differences in their amino acid sequences, subunit molecular weights, secondary tertiary and quaternary structures, thermal stability and catalytic mechanisms {179}. The type II DHQase monomer consists of a five-stranded parallel  $\beta$ -sheet core flanked by four  $\alpha$ - helices arranged with a similar overall topology to flavodoxin {179} .



**Figure 11.1 Conversion of Dehydroquinate to Dehydroshikimate in the presence of Typell DHQase**

The enzyme is a homododecamer and the subunits are arranged as a tetramer of trimers. Each subunit of type II DHQase has a molecular weight of approximately 16.5 kDa. The position of the active site was suggested by the cluster of conserved residues near the C-terminus of the  $\beta$ -strand {180}. The enzyme is most studied from the microorganisms *Streptomyces coelicolor*, *M. tuberculosis* and *Helicobacter pylori*. *Streptomyces coelicolor* was the first organism shown



to possess only a biosynthetic type II dehydroquinase rather than a type I dehydroquinase {181}. *Helicobacter pylori* is a Gram-negative bacterium and lives in stomach and duodenum of Human beings causing chronic gastritis, peptic ulcer and gastric adenocarcinoma, the type II dehydroquinase of *H.pylori* DHQase shows a strong identity to other members of type II dehydroquinase family {182}. Due to the absence of shikimate pathway in humans type II DHQase is considered to be an attractive target for the development of selective antimicrobial agents {142}. Several rounds of inhibitor development have been carried out leading to nanomolar inhibition levels against the enzyme {183}. One of the compounds 1,4,5-trihydroxy-3-(3-nitrophenyl)cyclohex-2-enecarboxylic acid (a 3-nitrophenyl derivative of dehydroquinone) with a  $K_i$  of 54.0nM is the most potent inhibitor reported against *M. tuberculosis* type II DHQase {184}. However most of them have performed poorly in whole cell tests (unpublished data). There is significant need to develop the pharmacokinetic properties of these molecules before these inhibitors can seriously be considered as potential antimicrobial agents for treating the diseases in humans.

## 11.1 Aims and objectives

1. Purification of different DHQases
2. Comparative characterization of the enzyme
3. Inhibition studies to be performed in collaboration with Dr Richard Payne, University of Sidney.

## 11.2 Expression and Purification of DHQases

All the DHQases plasmids were inserted in to the expression host BL21 (DE3) cells by using the transformation protocol. The *M. tuberculosis* DHQase (*MTB* DHQase) was purified by a three step process involving ion exchange chromatography, hydrophobic interaction chromatography and size exclusion chromatography as described in section 2.14. Fractions obtained after each purification step were tested for activity through standard DHQase enzyme assay and the fractions with maximum activity were pooled and loaded on to another column for further purification.

His tagged versions of *Streptomyces coelicolor* DHQase (SC DHQase), *Helicobacter pylori* DHQase (HP DHQase), *Campylobacter jejuni* DHQase (CJ DHQase) and *Candida albicans* DHQase (CA DHQase) were purified by using nickel purification method. The successive stages of purification of different DHQases in pictorial form are given below (Figure 12.2-7).

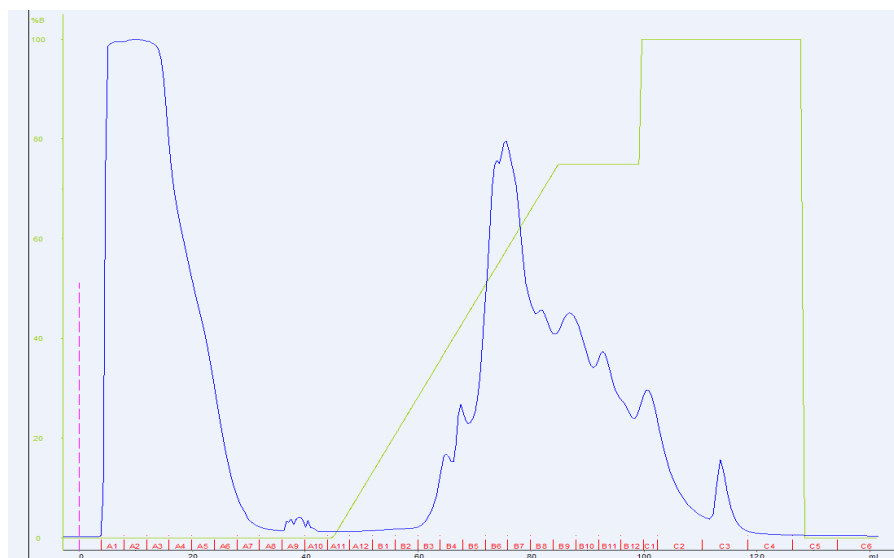
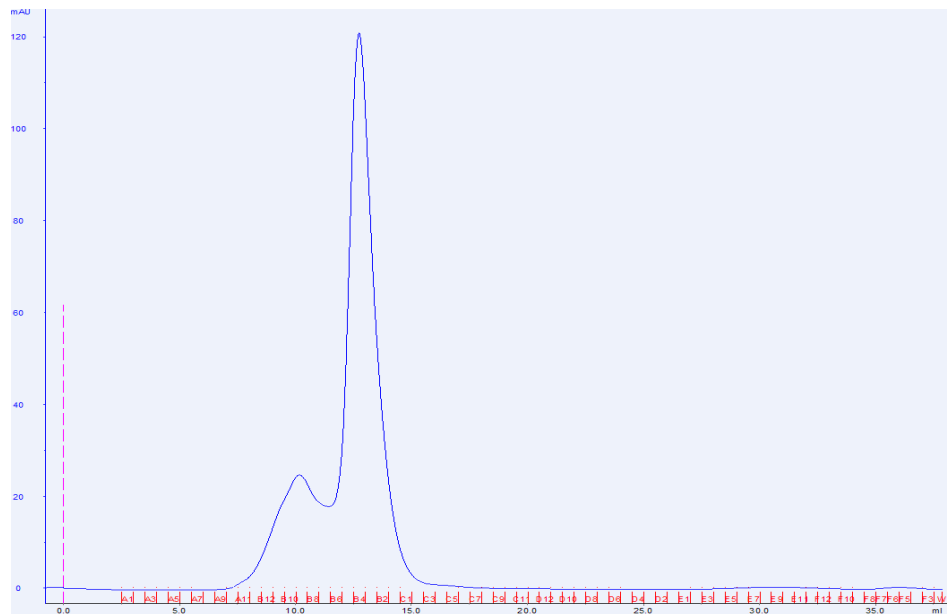


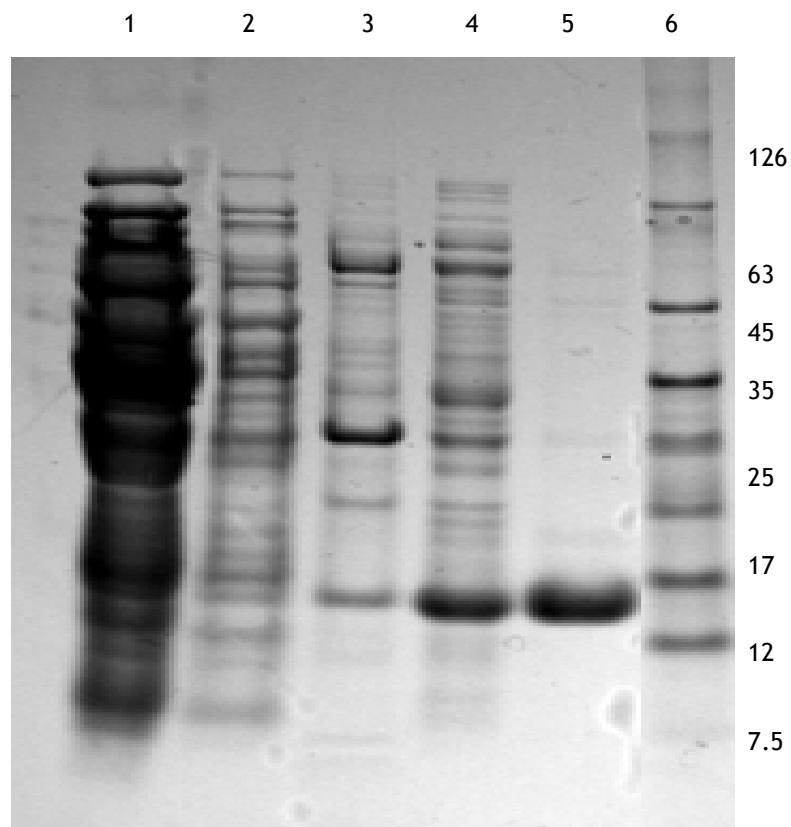
Figure 11.2 Chromatogram for *MTB* DHQase eluted on Q-sepharose Column (ion exchange chromatography)



Figure 11.3 Chromatogram for different fractions of *MTB* DHQase eluted on Phenyl sepharose column (HIC)



**Figure 11.4** Chromatogram image showing different fractions of *MTB* DHQase eluted on superdex 200 column



**Figure 11.5** SDS-PAGE analysis for *SC* DHQase after Ni-purification: 1= flow through, 2=70mM Imidazole wash, 3= 300mM Imidazole elute 4=500mM Imidazole elute, 5=700mM Imidazole elute, 6= molecular weight marker

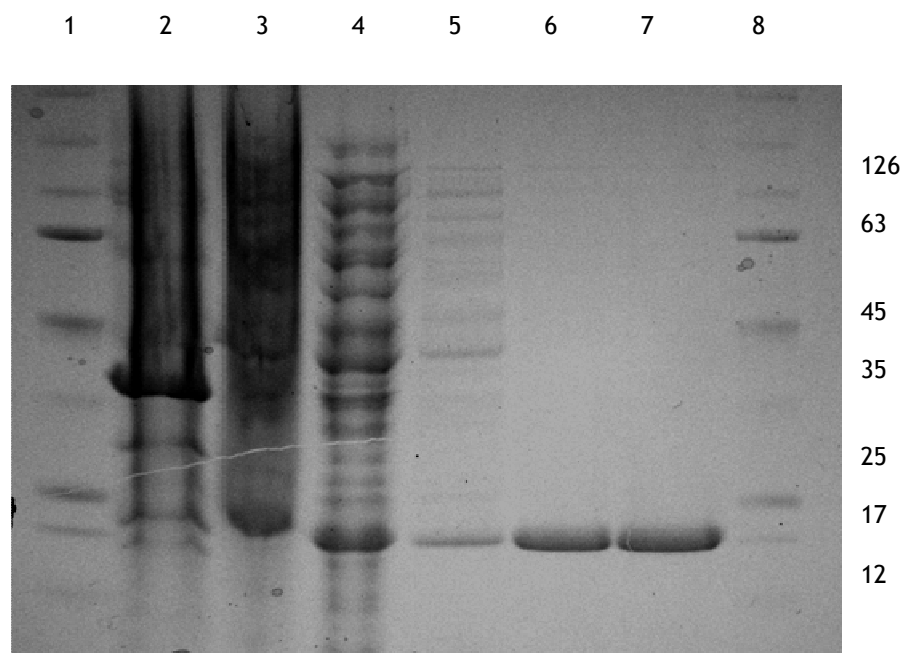


Figure 11.6 SDS-PAGE analysis for *HP* DHQase after Ni-purification: 1= molecular weight marker in Kda (Bio labs, NEW ENGLAND cat# P7708S), 2= cell pellet, 3= supernatant, 4=flow through, 5=70mM Imidazole wash, 6, 7=300mM Imidazole elute, 8= molecular weight marker

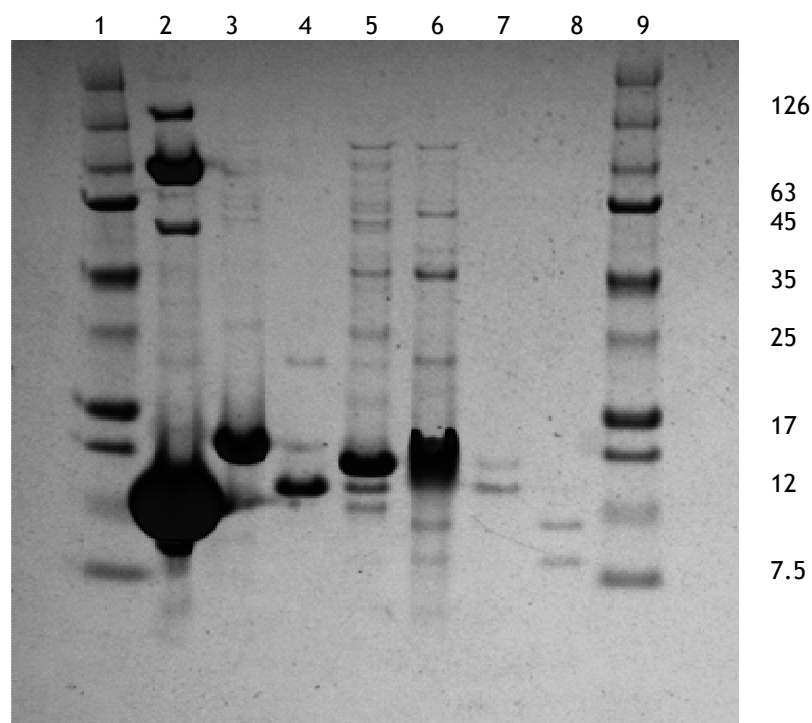


Figure 11.7 SDS-PAGE analysis for Purified DHQases: 1 = molecular weight marker in Kda (BioRad precision plus protein™ catalog no: 161-0373), 2 = *Bacillus subtilis* DHQase, 3 = *HP* DHQase, 4 = *MTB* DHQase, 5 = *SC* DHQase, 6 = *CJ* DHQase, 7 = *CA* DHQase, 8 = *MTB*, 9 = molecular weight marker

The purified samples of *MTB* DHQase, *SC* DHQase and *HP* DHQase were lyophilized and then sent to Dr. Richard Payne, University of Sidney, for enzyme inhibition assays.

## 11.3 Kinetic studies on DHQases

### 11.3.1 Enzyme Assays

The Standard assay to measure 3-dehydroquinase activity follows the formation of 3-dehydroshikimate. Dehydroquinone is added to the protein sample, and the rate of conversion of 3-dehydroquinone to 3-dehydroshikimate is monitored by the increase in absorbance at 234nm.

Enzyme kinetics were carried out to

1. Compare the  $K_m$  and  $K_{cat}$  values of DHQases from different microorganisms
2. Test the activity of the purified enzymes after lyophilization to confirm that freeze-drying the sample had caused no appreciable effect on the enzyme activity of the samples.
3. Compare the calculated values ( $K_m$ ,  $K_{cat}$ ) with previously reported values for different DHQases

All enzyme assays were carried out using Jasco V-550 dual beam spectrophotometer as described in section 2.19.1. Typically a 1mL path length quartz cell was used. Dehydroquinone was used as a substrate. Molar extinction coefficient for 3-dehydroquinone was measured to be  $12000 \text{ M}^{-1}\text{cm}^{-1}$ . All the assay measurements were made in triplicate by using 100mM Tris pH: 7.0 buffers. The assay data was used to calculate the  $K_m$  and  $K_{cat}$  values by using the Michalis-Menten equation and Line weaver-Burk plots (Table 11.1).

DHQase	$K_m$ ( $\mu\text{M}$ )		$K_{cat}$ ( $\text{s}^{-1}$ )		$K_{cat}/K_m$ ( $\text{s}^{-1} \text{M}^{-1}$ )	
	cal*	rep*	cal*	rep*	cal*	rep*
<i>HP</i>	200	205.8	0.91	0.94	$4.55 \times 10^3$	$4.55 \times 10^3$
<i>MTB</i>	32.8	23.8	3.3	5.2	$1.0 \times 10^5$	$2.18 \times 10^5$
<i>SC</i>	157.3	99.2	110.6	124.8	$7.03 \times 10^5$	$12.5 \times 10^5$
<i>CJ</i>	17.3	-	6.2	-	$3.58 \times 10^5$	-
<i>CA</i>	157.2	-	1.17	-	$7.44 \times 10^3$	-

**Table 11.1 Comparison of  $K_m$  and  $K_{cat}$  values of DHQases from different microorganism, all the enzyme assays were carried out at pH : 7.0 @ 25°C ( - = values not determined, cal = values calculated in the project work, rep\* = reported values from L.D.B Evans Thesis, 2003)**

The data obtained from enzyme kinetics showed that *CJ* DHQase has the lowest  $K_m$  value, while *CA* DHQase has a  $K_m$  values similar to *SC* DHQase. The  $K_m$  values of DHQases from different microorganisms indicated that the substrate binds most tightly to the *CJ* DHQase and most weakly to *HP* DHQase. The binding affinity in descending order is given as

$$CJ \text{ DHQase} > MTB \text{ DHQase} > CA \text{ DHQase} = SC \text{ DHQase} > HP \text{ DHQase}$$

The  $K_{cat}$  value was highest for *SC* DHQase and lowest for *CA* DHQase. The  $K_{cat}$  values for different DHQases in descending order are given as

$$SC \text{ DHQase} > CJ \text{ DHQase} > MTB \text{ DHQase} > CA \text{ DHQase} > HP \text{ DHQase}$$

The kinetic data indicated no loss in the enzyme activity of the samples after lyophilization. There was a small difference observed between calculated and reported values which could be due to slight changes in pH or temperature or a high concentration of NaCl in the lyophilized samples.

## 11.4 Crystal structure of *C.jejuni* DHQase

*CJ* DHQase was over expressed and purified by Sabine Schwartz (Project student B.Sc (Hons)). To obtain the crystal structure of the enzyme substrate complex (ESC) crystallization trials were carried out at a range of conditions. The best

conditions to grow crystals were identified under the microscope. Initially the crystals were obtained from the crystallization trials of M-Screen (M1-M114) at M88 and M75. These conditions were then slightly optimized to get better crystals. The most promising individual large crystals were obtained at [0.1M MOPS (pH: 6.5), 15 % ME2K PEG, 0.2M  $\text{MgCl}_2$ ].

#### 11.4.1 *Crystallographic Data collection and processing*

X-ray diffraction data were collected for *CJ* DHQase in house on a Mar345dtb image plate & Rigaku Micromax 007 X-ray generator with Osmic mirrors. The crystal was briefly soaked with 2mM DHQ in the crystallisation mother liquor along with 30% glycerol and then flash frozen at 100K in a stream of gaseous nitrogen using an Oxford cryostream. 180 degrees of data was collected from a single *CJ* DHQase crystal. The data was indexed as face-centered cube with unit cell dimensions  $a = b = c = 127.53\text{\AA}$  by using the program IMosflm. This meant that the data was highly redundant and a unique dataset required only 30 degrees of data. Typical diffraction image is shown in figure 11.8. Altogether 360 frames were collected with an oscillation angle of  $0.5^\circ$ . Data collection and processing statistics for the first 60 degrees are given in table 11.2.

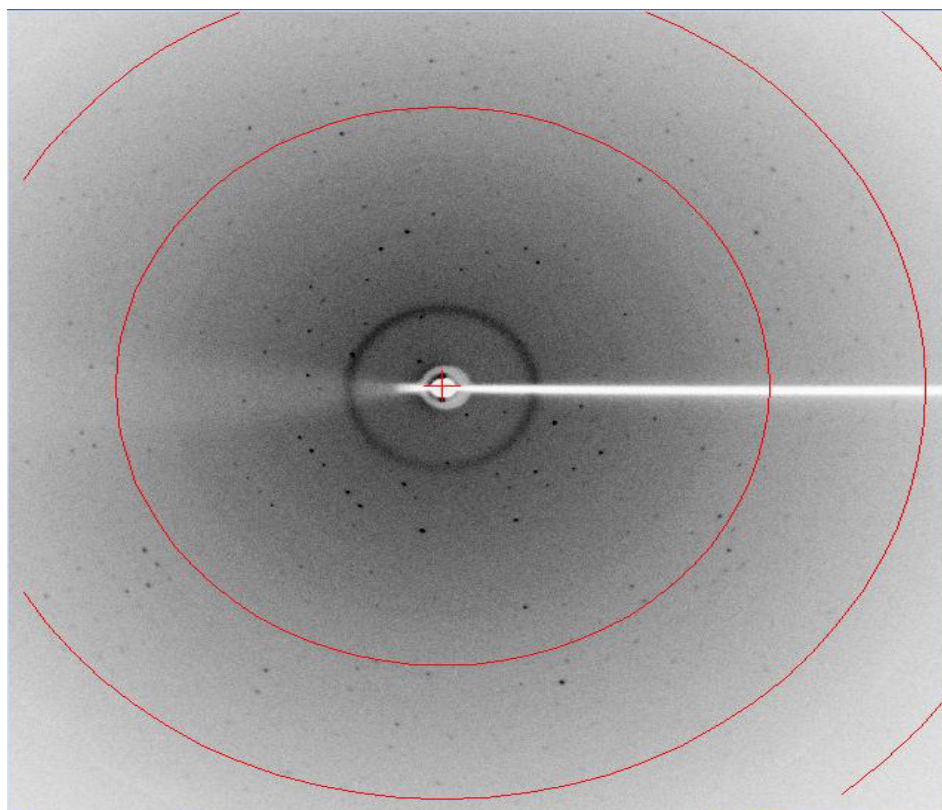


Figure 11.8 A typical diffraction image obtained from a crystal of *CJ* DHQase

The crystal was very mosaic with a refined mosaicity of  $1.5^\circ$ . The diffraction was strong around  $2.8\text{\AA}$  resolution, but extended weakly to  $2.4\text{\AA}$  resolution. The data was complete with an average multiplicity of 6.1 and  $R_{\text{merge}}$  of 21.9 %. The X-ray data processing and subsequent calculations were carried out using programs from the CCP4 suite of programs [185]. The structure was solved by using the *HP* DHQase monomer coordinates (PDB: 2C4V). The search model (PDB: 2C4V) was employed by using the program PHASER. The program gave a unique solution with rotation function Z-score of 5.7 and a translation function Z-score of 20.5 and a final log likelihood gain of 294. The structure was refined and adjusted by iterative cycles of refinement using the program REFMAC5. The model was built by using the program COOT®. The representative electron density for the structure is shown in figure 11.9. Interestingly the crystal structure had the product (3-dehydroshikimate) in the active site. Figure 11.10 shows the final electron density around the product (Dehydroshikimate, DHS) in the active site of the enzyme. The final structure had good geometry as assessed by PROCHECK. The structure gave an excellent Ramachandron plot with 91% of amino acid residues within the most favorable region and none in disallowed regions (Figure 11.11). The refined structure had an R-factor of 19.8 % and a free R-factor of 26.5 %.



---

**Data collection statistics**


---

<b>Data set</b>	<b><i>CJ</i> DHQase and 3-dehydroshikimate</b>
Space group	F23
Unit cell	a=b=c =127.53Å
wavelength	1.5418Å
Crystal to Detector distance(mm)	220
Resolution range/ High Resolution	38.47-2.36Å (2.49-2.36Å)
completeness	97.4% (82.6%)
Unique reflections	6971
Merging R-factor	0.219
Wilson B-value	34.0
Multiplicity	6.1 (5.1)
Mean [I/Sd(I)]	7.7 (1.7)
<b>Refinement statistics</b>	
Resolution range	36.8-2.37Å
Overall R-factor	0.1986
Free R-factor	0.2649
Over all B-value	43.4
RMS bond length deviation	0.0142Å
RMS bond angle deviation	1.2637Å

---

**Table 11.2 X-ray crystallographic data for *CJ* DHQase**

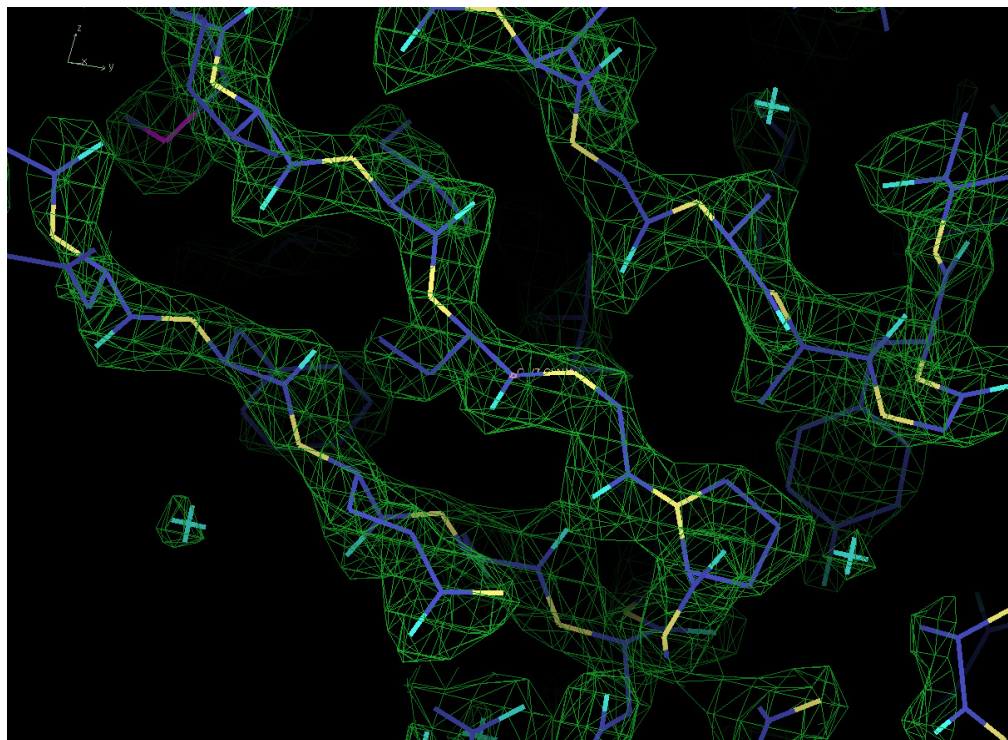


Figure 11.9 Electron density map for *CJ* DHQase at a resolution of 2.4Å (figure generated by Coot®)

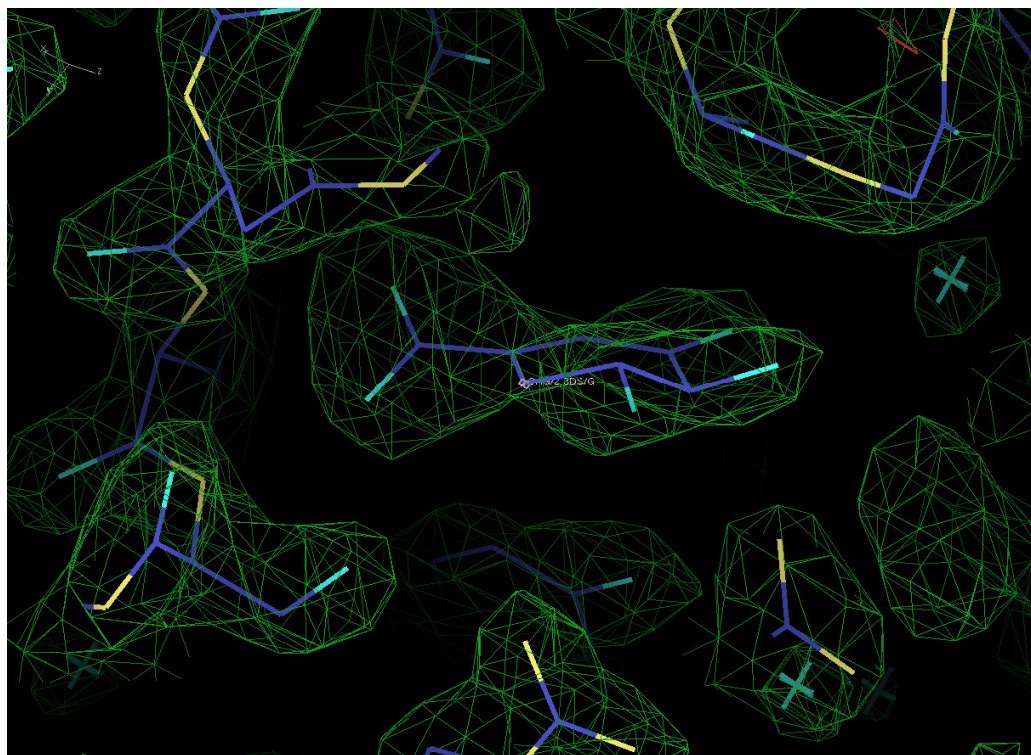
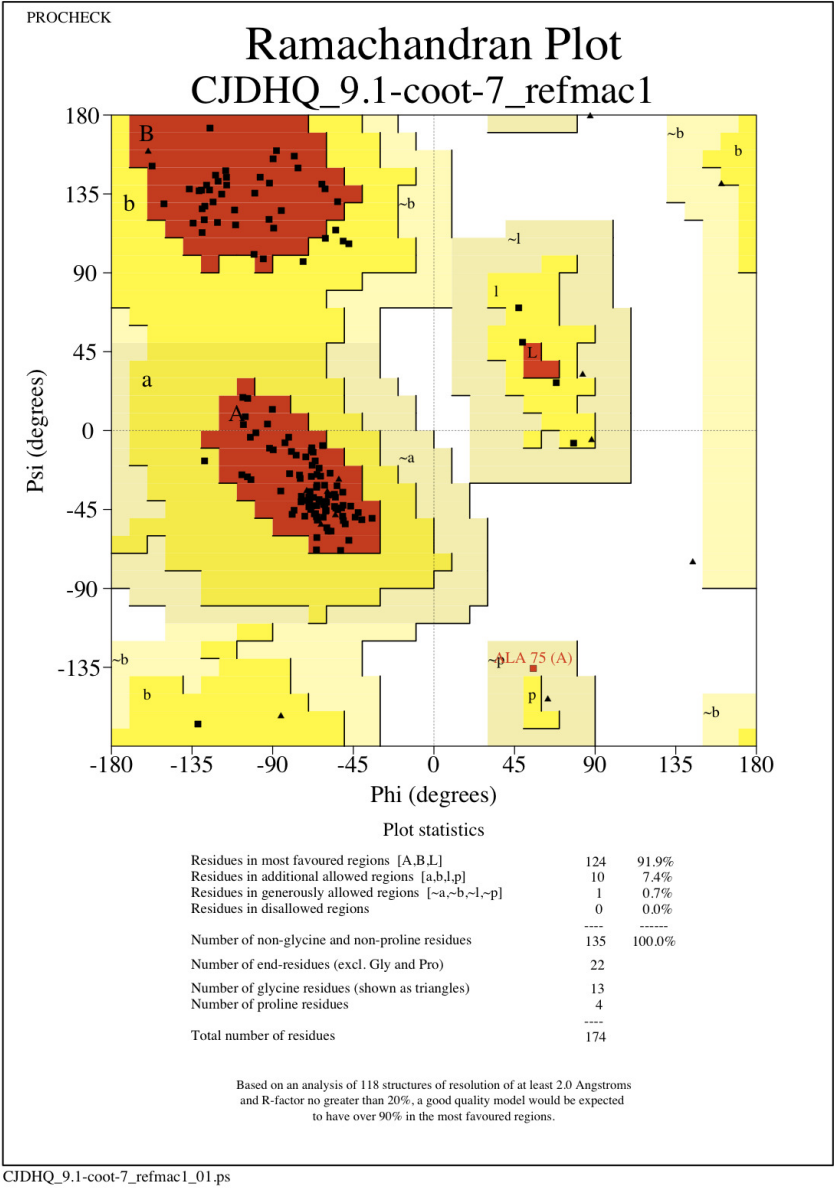


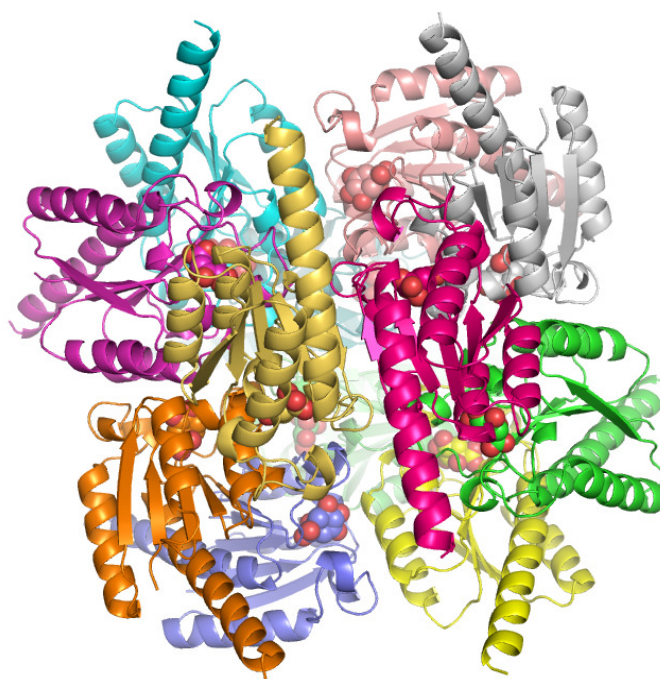
Figure 11.10 Electron density map for 3-Dehydro shikimate in the active site of *CJ* DHQase at a resolution of 2.4Å (figure generated by Coot®)



**Figure 11.11** Ramachandran plot for crystal structure of *CJ* DHQase obtained by using PROCHECK

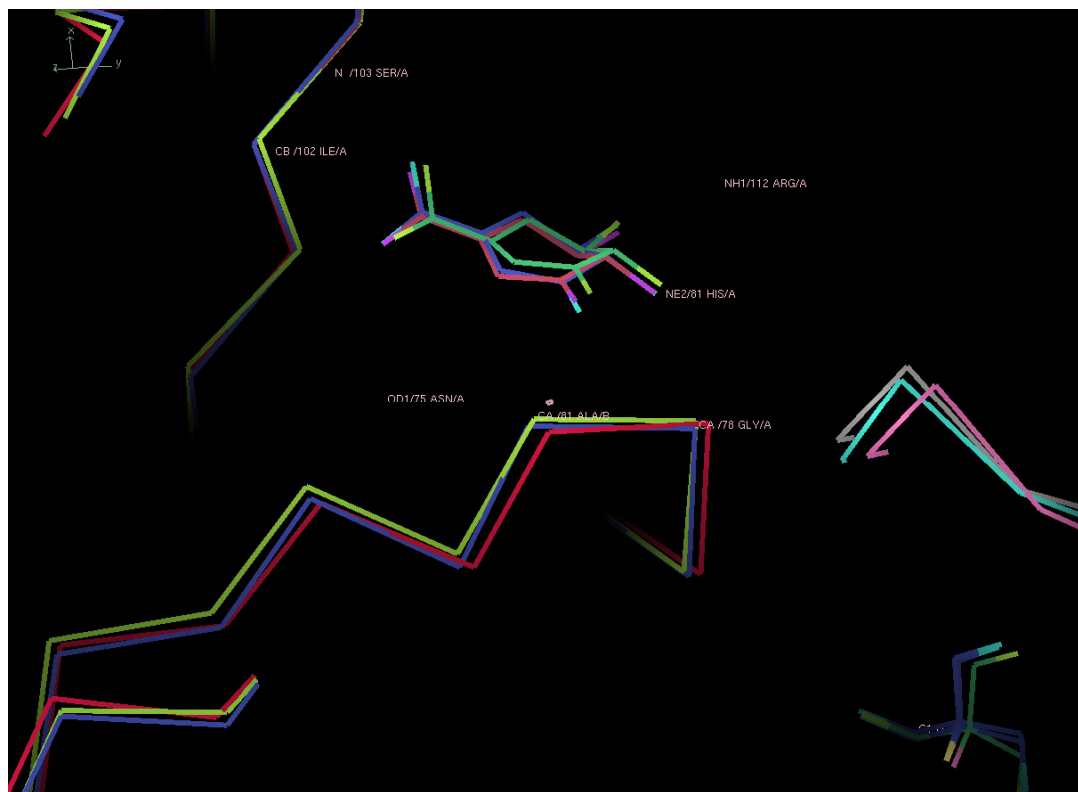
### 11.4.2 Structure analysis

The visualization of the final structure in Coot indicated that DHS formed good H-bonds with the surrounding active site residues. The catalytic reaction product 3-dehydroshikimate was observed in the subunits (the substrate was converted into product during crystallization). The binding of DHS is stabilized by hydrogen-bonding interactions with the backbone amides of Ile102 and Ser103, with the side-chain atoms of Arg19, Asn75, His81, Arg112 and Asp88\* (\* indicates residues from an adjacent subunit) and by hydrophobic interactions of its cyclohexene moiety with the sidechain atoms of Tyr24, His101, Leu13, Val105, Arg112 and the C alphas of Ala77 and Gly78. The dodecameric structure shows that DHS is located at the interface of two subunits (Figure 11.12).



**Figure 11.12** Ribbon Diagram for the dodecameric structure of *CJ* DHQase along with DHS (CPK model) in the active site of the protein (Figure generated by Pymol).

A comparative study of the x-ray structures of DHQases including *CJ* DHQase, *MTB* DHQase and *SC* DHQase demonstrated that *MTB* DHQase and *CJ* DHQase contain a Glycine 78 residue while the *SC* DHQase has an Alanine at this position in the active site. The superimposed structures through Coot® indicated that DHS fits well in the active site of *CJ* DHQase & *MTB* DHQase while in case of *SC* DHQase it has a different conformation. The different conformation of DHS in *SC* DHQase can be due to the presence of Alanine residue which is replaced by a Gly78 in *CJ* DHQase & *MTB* DHQase (Figure 11.13).



**Figure 11.13** Superimposed C-Alphas structures of *C.jejuni*, *M.tuberculosis* & *S.coelicolor*. DHS structure in blue and red color is from the structures of *CJ* & *MTB* while the DHS in green color is from the structure of *SC* DHQase.

The sequence analysis of active site of different DHQases shows that this Glycine is unique only in *CJ* DHQase & *MTB* DHQase but is replaced by an Alanine in other DHQases like *SC* DHQase, *HP* DHQase & *CA* DHQase. Interestingly the  $K_m$  value is the lowest for the former DHQases but is significantly high for the latter DHQases (Table 11.3).

DHQase	Active Site Sequence	K <sub>m</sub> (μM)
<i>CJ</i>	N A A <span style="border: 1px solid black;">G</span> Y T H	17.3
<i>MTB</i>	N A G <span style="border: 1px solid black;">G</span> L T H	32.8
<i>SC</i>	N P A <span style="border: 1px solid black;">A</span> Y S H	157.3
<i>CA</i>	N A G <span style="border: 1px solid black;">A</span> Y T H	157.0
<i>HP</i>	N P G <span style="border: 1px solid black;">A</span> F S H	200.0

**Table 11.3 Comparative account of amino acid sequences of different DHQases and their respective K<sub>m</sub> values (the residue numbering for *CJ* DHQase is N75, A76, A77, G78, Y79, T80, H81)**

The presence of a Gly residue within the active site provides more room for the surrounding active site residues to form close contacts with DHS but in case of *SC* DHQase the presence of Ala at the same position moves the same residues a little away from DHS. Further study of the superimposed structures revealed that the presence of Alanine in place of Gly78 in the structure of *S.coelicolor* causes a slight movement of the symmetry related Asp 88<sup>#</sup> and other adjacent residues away from the DHS. A lower K<sub>m</sub> value for *CJ* DHQase & *MTB* DHQase is consistent to the presence of this glycine residue which is lacking in other DHQases like *SC* DHQase, *HP* DHQase and *CA* DHQase (Table 12.3).

To analyze the interactions between DHS and different DHQases, the H-bond distances and angles were measured for *CJ* DHQase, *MTB* DHQase & *SC* DHQase by using Coot®. When compared to *MTB* & *SC* DHQase it appeared that *CJ* DHQase formed better H-bond distances and angles towards the carboxyl moiety of DHS (Table 12.4). Further towards the six membered ring of DHS the H-bond distances were relatively better between His81 and *CJ* DHQase than *MTB* & *SC* DHQase. In case of Arg112 an opposite trend was observed where the H-bonding was better for *SC* & *MTB* DHQase than *CJ* DHQase. The over all H-bond data showed that in general the interactions between DHS and the active site residues were better for *CJ* DHQase than *MTB* & *SC* DHQase.



Residue number	bonding atoms	<i>CJ</i> DHQase	<i>MTB</i> DHQase	<i>SC</i> DHQase
Arg17	NH1...O3—C3	2.83 (125.09)	2.94 (134.1)	—
Ala77 <sup>^</sup>	CB...O3—C3	3.13 (112.35)	—	3.21 (94.50)
Gly78	N...OH4—C4	3.13 (117.89)	3.11 (125.0)	—
Asp88 <sup>#</sup>	OD2...OH4—C4	2.57 (111.45)	2.47 (119.3)	2.68 (116.42)
Arg112	NH1...OH4—C4	2.96 (85.84)	3.22 (83.64)	2.98 (140.56)
	NH1...OH5—C5	2.93 (108.80)	2.83 (115.7)	2.81 (115.64)
His81	NE2...OH5—C5	2.63 (115.42)	2.68 (119.8)	2.77 (113.34)
Ile102	N...2O7—C7	2.78 (105.48)	2.90 (100.0)	3.08 (100.14)
Ser103	OG...1O7—C7	2.61 (143.31)	2.84 (138.8)	2.59 (154.22)
	N...1O7—C7	2.66 (104.71)	2.80 (134.9)	2.88 (140.66)
Asn75	ND2...2O7—C7	2.99 (158.69)	3.15 (148.5)	3.03 (150.83)

**Table11.4 Comparative data for H-bond lengtha and angles between DHS and active site residues of different DHQases. <sup>^</sup> = short contact, <sup>#</sup> = contact from the adjacent subunit, — = residue absent, the distances are in Å and corresponding angles are given in parenthesis.**

## 11.5 Future work

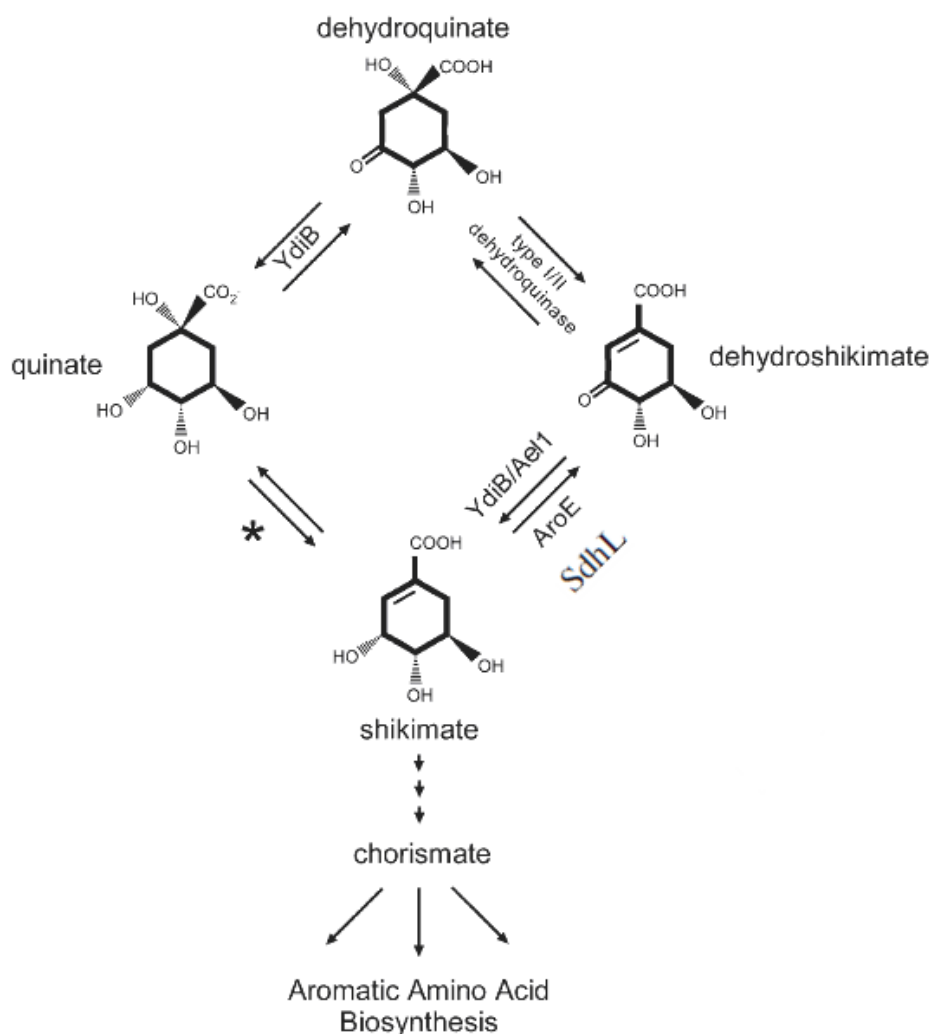
Based on the low  $K_m$  values obtained for different DHQases and the sequence analysis it would be sensible to mutate the Glycine residue with Alanine in *CJ* DHQase and *MTB* DHQase and then carrying out the enzyme assay on mutated form of these enzymes for measuring the  $K_m$  values. To Further confirm these studies a reverse study can be carried out by mutating the Alanine residue to Glycine in the active site of DHQases with high  $K_m$  values like *SC* DHQase, *HP* DHQase & *CA* DHQase. These experiments could reveal in finding out the crucial role of Glycine (Gly78) in the active site of the DHQases. Obtaining the X-ray structure of these mutated versions would be helpful in exploring the nature of interactions of surrounding residues between themselves and with the true ligand in the active site of the enzymes.

## 12. *HISdhL (Haemophilus Influenzae SdhL)*

The shikimate dehydrogenase (SDH) represents a widely distributed protein enzyme family and performs essential roles in the secondary metabolism of microbes and plants. The shikimate pathway occupies a central position for aromatic biosynthesis in microbes and plants but is absent in humans and other higher animals. The absence of the shikimate pathway in animals makes it an ideal target for herbicide and for the anti-microbial drug design. Recently the shikimate pathway was identified in apicomplexan parasites, including *Toxoplasma gondii* and *Plasmodium falciparum*, which has renewed interest in better understanding of the enzymes in the pathway. To date 5 classes of shikimate dehydrogenases have been identified and characterized, namely AroE, YdiB, YdiB2, SdhL, and RifI {186, 187}.

The combination of biochemical, phylogenetic, and genomic approaches has revealed a broad extent functional diversity in the SDH enzyme superfamily. All the classes in the family show clear biochemical and functional differences ranging from amino acid biosynthesis to antibiotic production {186}. YdiB is a bifunctional enzyme that catalyzes the reversible reductions of dehydroquinate to quinate and dehydroshikimate to shikimate (Figure 12.1) in the presence of either NADH or NADPH as a co-factor {187}. Another analogue of the SDH family is *HISdhL* whose kinetic properties are remarkably different from those of AroE and YdiB. In comparison to YdiB, *HISdhL* catalyzes the oxidation of shikimate to dehydroshikimate but not quinate. *HISdhL* turnover rate for the oxidation of shikimate is 1000-fold lower as compared with that of AroE. The kinetic properties of *HISdhL* suggest that both shikimate and quinate are not its preferred substrates {186, 187}.





**Figure 12.1** Role of different shikimate dehydrogenases in the shikimate/quinate pathway, \* = quinate hydrolyase, image modified from [186]

The previous kinetic studies, carried out in our group (L.D.B Evans PhD Thesis University of Glasgow) have revealed that *HISdhL* and *HIAroE* are both  $\text{NAD}^+/\text{NADP}^+$  dependant dehydrogenases. Both are able to utilize shikimate in the presence of co-factors ( $\text{NAD}^+$ ,  $\text{NADP}^+$ ), except *HIAroE* which is able to utilize even quinate in the presence of both co-factors. The data shows that *HIAroE* binds more tightly to  $\text{NADP}^+$  ( $K_{\text{NADP}} = 20\mu\text{M}$ ) than *HISdhL* ( $K_{\text{NADP}} = 380\mu\text{M}$ ) but weaker to  $\text{NAD}^+$  ( $K_{\text{NAD}} = 8500\mu\text{M}$ ) in comparison to *HISdhL* ( $K_{\text{NAD}} = 1400\mu\text{M}$ ). Further the data shows that *HIAroE* binds more tightly to shikimate ( $K_{\text{m}} = 0.02\text{mM}$ ) than *HISdhL* ( $K_{\text{m}} = 0.38\text{mM}$ ), similarly the  $K_{\text{cat}}$  values for *HIAroE* are much higher than *HISdhL* i.e.; 194 and 19 respectively. Based on interrogation of the active site of *HISdhL* by our group, it was proposed that a potential substrate could be

Mannose-6-phosphate. The kinetic studies carried out for *HISdhL* with different substrates (Table 12.1) confirmed that mannose-6-phosphate is indeed a much better substrate of *HISdhL* with a  $K_m$  values of 56  $\mu\text{M}$  in comparison to shikimate with a  $K_m$  value of 30500  $\mu\text{M}$  (unpublished data). The specificity factor ( $K_{\text{cat}}/K_m$ ) for mannose-6-phosphate is not particularly high suggesting that it may not be the true substrate for the enzyme, however there is a significant difference in kinetic parameters compared to the only other identified substrate i.e. shikimate.

Substrate	$K_{\text{cat}}$ ( $\text{s}^{-1}$ )	$K_m$ ( $\mu\text{M}$ )	$K_{\text{cat}}/K_m$ ( $\text{M}^{-1}\text{s}^{-1}$ )
Shikimate	4.5	30500	148
Quinate	—	—	—
Mannose	<0.005	—	—
Glucose	<0.005	—	—
Mannose-6-phosphate	0.47	56	8393
Glucose-6-phosphate	<0.005	—	—

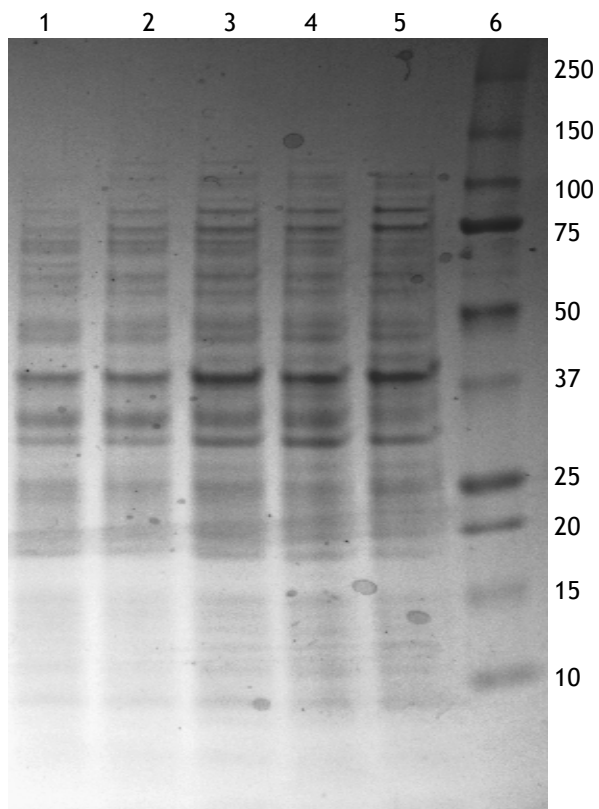
**Table 12.1** Kinetic data for *HISdhL* against different potential substrates (— = Not obtainable, unpublished data)

As till date no structure of *HISdhL* has been determined with the co-factor and/or substrate, the determination of the crystal structure of *HISdhL* along with mannose-6-phosphate would significantly help in understanding the true function of this class of enzymes.

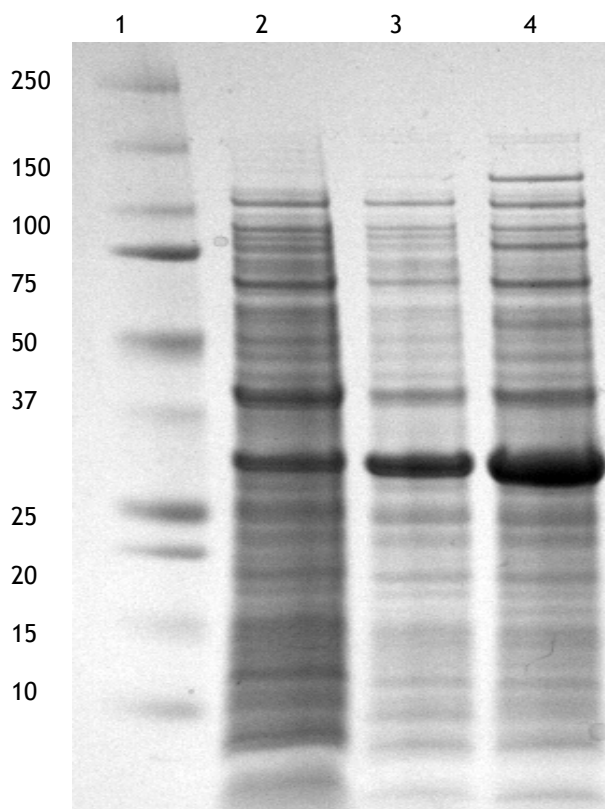
## 12.1 Expression and purification

*HISdhL* is a protein comprising of 277 amino acids with a molecular weight of 30.7 Kda. The Non-His tagged version of the protein has been previously purified by using a different construct (L.B Evans PhD thesis, 2004 University of Glasgow). Recently the gene for *HISdhL* was cloned from the PTb361 into to another vector to form a new His tagged vector called PTbL2. The aim of attaching His-tag to the enzyme was easy and short cut purification of the protein. The PTbL2

plasmid was successfully transformed into BL21 (DE3) competent cells. The over-expression of the protein was carried out in LB media at 37°C however, the yield was very low (Figure 12.2). In order to increase the over-expression of protein the cultures were first grown at 37°C and as the O.D<sub>600</sub> reached to 0.6, the cultures were put in slushy ice for 5 minutes and then induced with 1mM IPTG and left for overnight induction at 13°C. The induction at 13°C significantly improved the over-expression (Figure 12.3). Precision plus protein™ BioRad, cat# 161-0373 molecular weight marker was used in all the SDS-PAGE analysis.

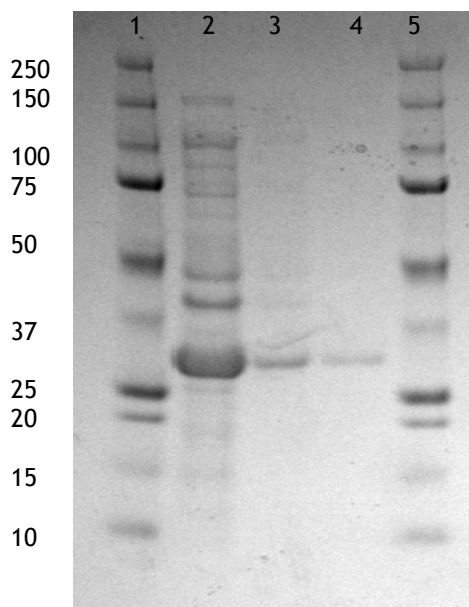


**Figure 12.2 SDS-PAGE analysis for *H/SdhL* expressed in LB media at 37°C 1=molecular weight marker, 2= uninduced sample, 3=overnight induced sample, 4= overnight induced sample at 13°C)**



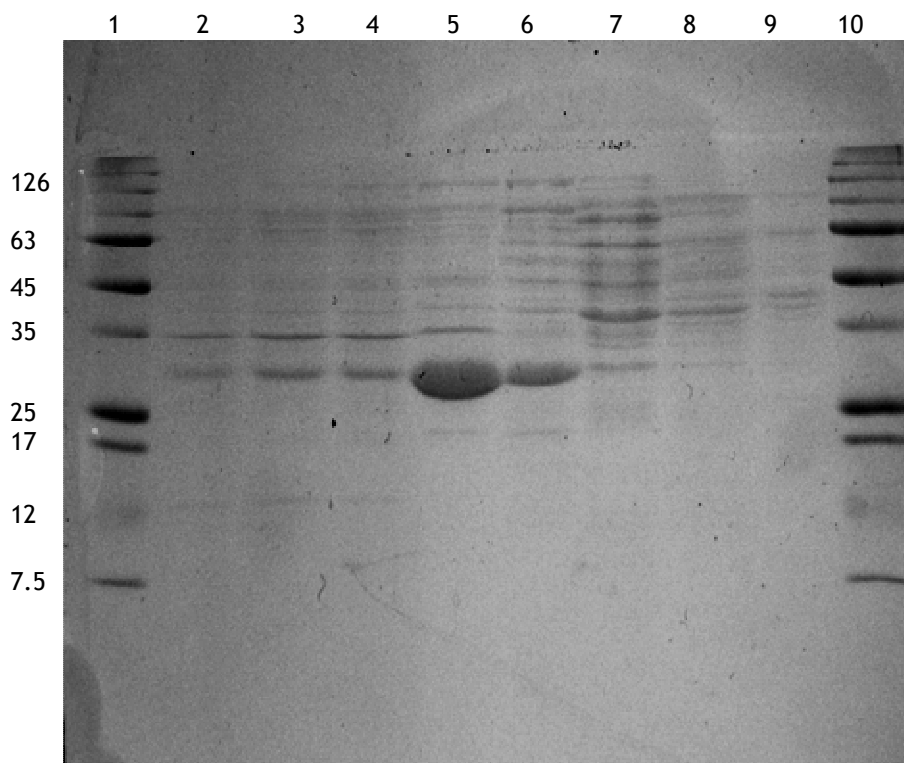
**Figure 12.3 .SDS-PAGE analysis for *H/SdhL* expressed in LB media at 37°C 1=molecular weight marker, 2= uninduced sample, 3=overnight induced sample, 4= overnight induced sample at 13°C)**

As the His-tagged version of the protein did not bind very tightly to the Ni-column (Figure 12.4.) which posed some problems during the purification of the protein. Attempts were made to design a step wise plan for the purification of the protein to get the maximum yield.



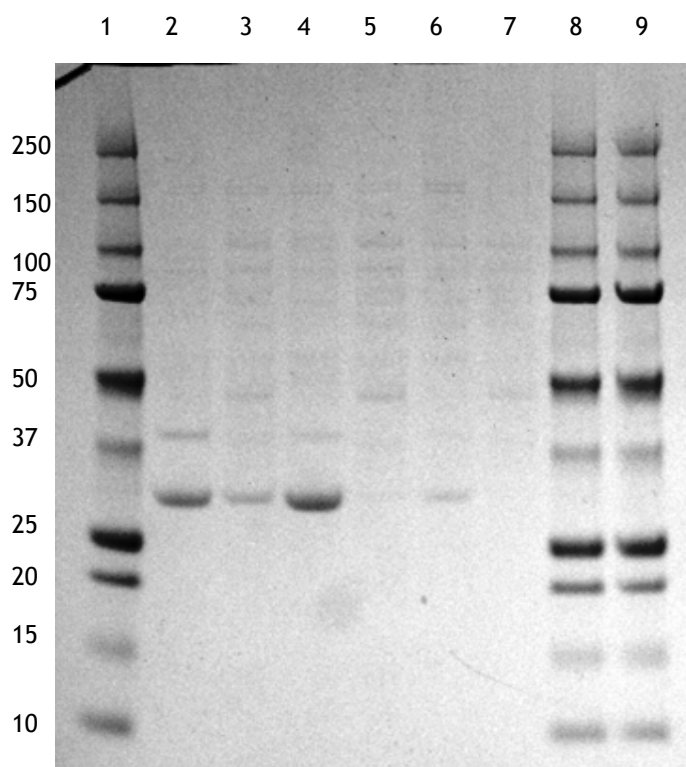
**Figure 12.4 SDS-PAGE analysis for *H/SdhL* after Nickel Purification: 1=marker, 2=flow through, 3=75mM Imidazole wash, 4=300mM Imidazole elute, 5=marker**

The over-expressed N-terminal His tagged protein was purified by first passing through Ni-column and then subjected to sulphate cuts. The sulphate cuts (5-40%) were performed on the flow through from Ni-purification by adding powdered ammonium sulphate to the flow through with a magnetic stirrer at 4°C on a magnetic plate and then centrifuged at 20000xg for 20 minutes at 4°C. The precipitate (protein) obtained was redissolved in the buffer (50mM Tris, 500mM NaCl, pH: 7.5) and to the supernatant additional powdered ammonium sulphate was added to increase the %age of ammonium sulphate and was then processed in the same way. The samples obtained were analyzed by running on a SDS-PAGE gel (Figure 12.5)



**Figure 12.5 SDS-PAGE analysis for *His*SdhL after sulphate cuts 1 = marker, 2 = precipitate after 5% cut, 3 = precipitate after 10% cut, 4 = precipitate after 15% cut, 5 = precipitate after 20% cut, 6 = precipitate after 25% cut, 7 = precipitate after 30% cut, 8 = precipitate after 35% cut, 9 = precipitate after 40% cut, 10 = marker**

The gel image showed that the maximum amount of protein is precipitated out at 20-25% of ammonium sulphate (Figure 12.6). The protein samples with 20% and 25% ammonium sulphate were pooled. The excess amount of ammonium sulphate was removed from the protein sample by buffer exchanging and concentrating down by using vivaspin in 100mM tris, 1mM DTT pH: 7.4 buffer.



**Figure 12.6** SDS-PAGE analysis for *HISdhL* after sulphate cuts 1 = marker, 2 = precipitate after 20% cut, 3 = supernatant from 20% cut, 4 = precipitate from 25%cut, 5 = supernatant from 25% cut, 6 = precipitate from 30% cut, 7 = supernatant from 30%cut, 8 = marker, 9 = marker

The concentrated protein sample was then loaded on phenyl sepharose column (HIC). The protein was eluted from the column (Figure 12.7) with a gradient of 500-0 mM NaCl of Buffer B (100mM Tris, 500mM NaCl, 1mM DTT, pH: 7.4) against Buffer A (100mM Tris, 1mM DTT, pH: 7.4).

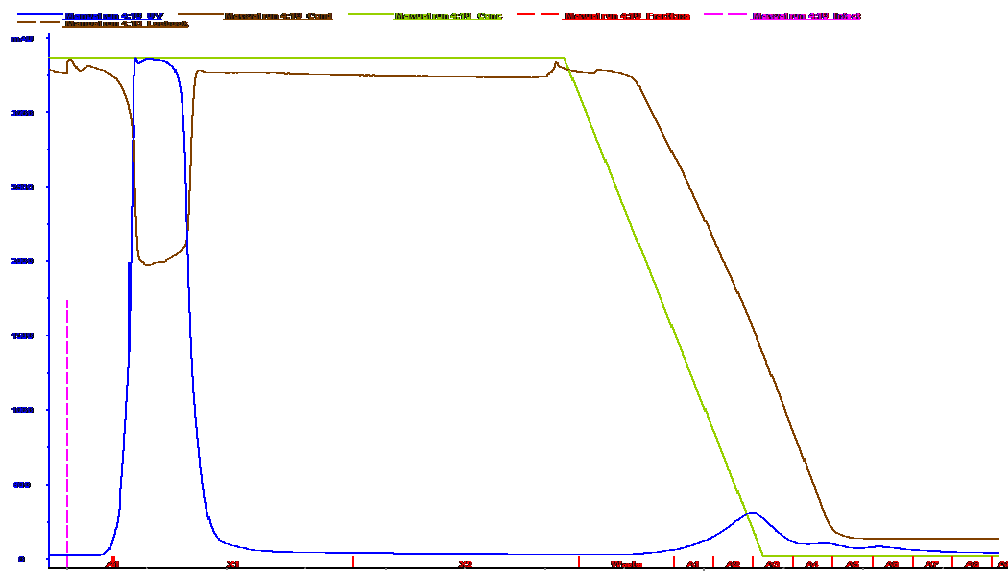


Figure 12.7 Chromatogram image for *HISdhL* after passing through phenyl sepharose column (HIC)

The SDS-PAGE analysis on the fractions obtained after HIC showed that the protein did not bind to the column. Therefore to further purify the protein the pooled fractions were loaded on SP sepharose column (ion exchange chromatography), as described in section 2.14.1. The fractions (Figure 12.8) obtained from SP sepharose column were subjected to SDS-PAGE analysis for checking the purity of the samples (Figure 12.9). The pure fractions were concentrated down to the desired values and further used for setting up co-crystallizations.

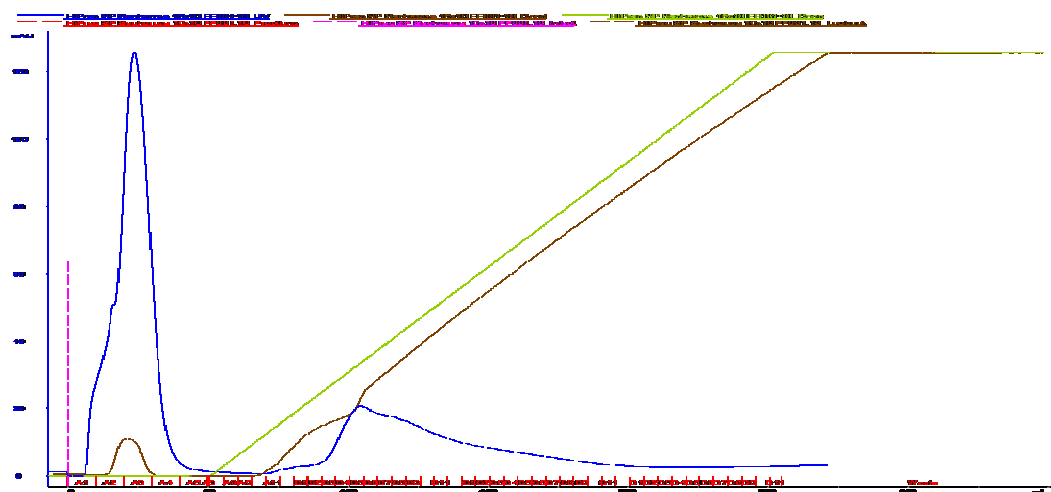
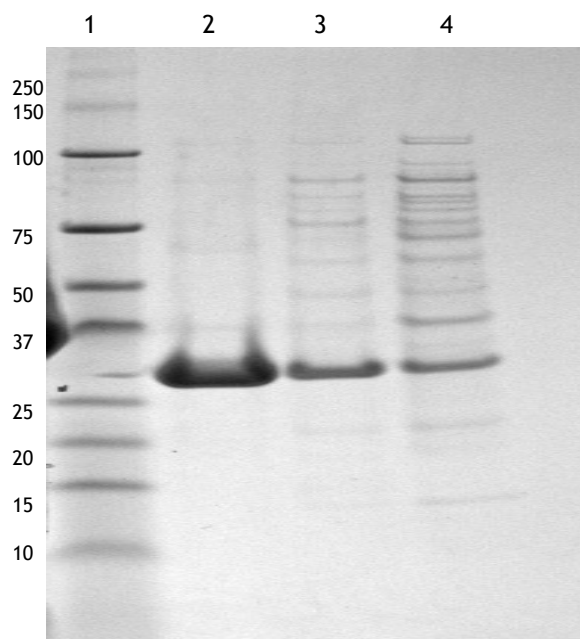


Figure 12.8 Chromatogram image for *HISdhL* after passing through SP sepharose column (Ion exchange chromatography)



**Figure 12.9** SDS-PAGE analysis for *H/SdhL* after passing through sp sepharose column 1 = marker, 2 = A2, 3=A3, 4 = B5

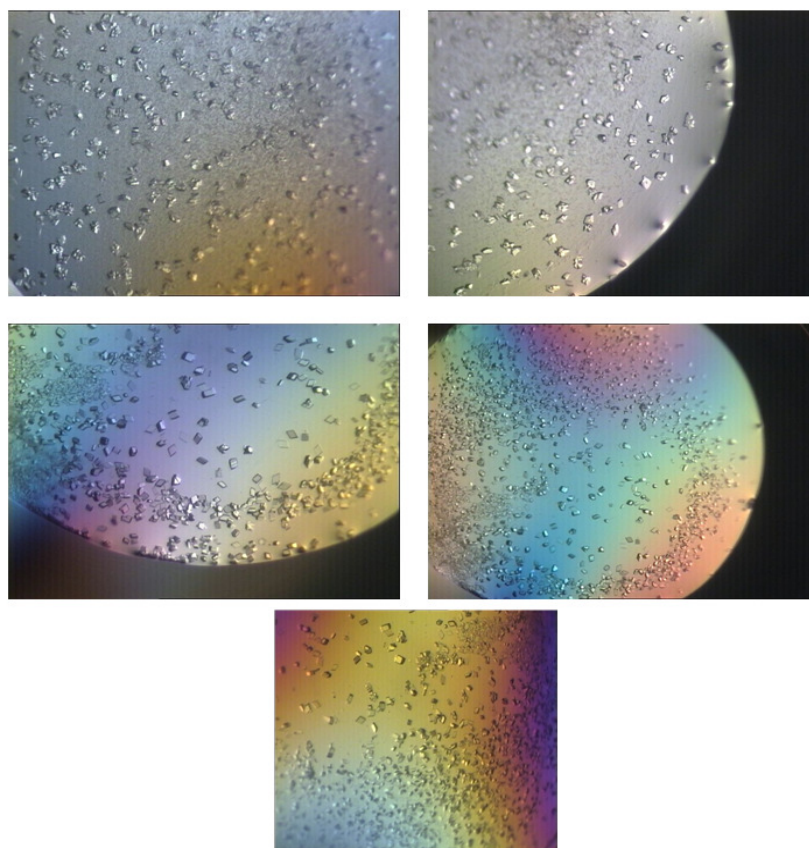
## 12.2 Crystallization of *H/SdhL*

Through the enzyme assays carried out previously it has been confirmed that mannose-6-phosphate is a substrate of *H/SdhL* and the co-factor for *H/SdhL* is  $\text{NADP}^+$  and  $\text{NAD}^+$ . The objective of co-crystallizing the protein was to obtain a crystal structure of the protein along with the potential ligand to help in understanding the active site interactions with the ligand and thereby assisting in knowing the functional mechanism of the protein. Crystallizations were performed by using sitting-drop vapour-diffusion method. Various crystallization conditions were carried out to co-crystallize the protein along with the substrate and co-factor. After initial difficulties in generating crystals, some small crystals were obtained at a number of conditions. The conditions at which best crystals grew are given in (Table 12.2). In order to obtain better crystals the crystallization conditions were further optimized and individual crystals were obtained (Figure 12.10). To prevent damage to the crystals, cryoprotector was prepared for each condition at which the crystals were obtained. Crystals were dipped in the respective cryoprotector containing 30% glycerol prior to diffraction. The best crystals were selected for diffraction.



C.No	0.1M Buffer/ pH	%age of PEG 3350	Crystals of
1	Na-Acetate/5.4	9	Protein+ mannose-6-phosphate (substrate)
2	Citrate/6.6	18	Protein + NADP <sup>+</sup> (co-factor)
3	Citrate/6.3	19	Protein only
4	Citrate/6.6	17	Protein+ co-factor
5	Citrate/5.88	20	Protein only
6	Citrate/5.88	17	Protein+ substrate+ co-factor
7	Citrate/5.88	15	Protein+ co-factor
8	Citrate/5.88	16.5	Protein only
9	Citrate/5.88	18	Protein+ co-factor
10	Citrate/5.88	16	Protein only

**Table 12.2 Crystallization conditions for individual crystals of HISdhL along with the substrate and co-factor**



**Figure 12.10 Images of *HISdhL* crystals obtained from conditions No 3, 4, 6, 7 & 9**

It was observed that all the crystals diffracted up to a maximum resolution of 3.0Å and no more than that. To get the resolvable electron density for substrate and co-factor it was tried to get the diffraction images at a higher resolution below 3.0Å. After a number of attempts, the crystals diffracted to a maximum resolution of 3.0Å (Figure 12.11) and no diffraction pattern was visible at a higher resolution.

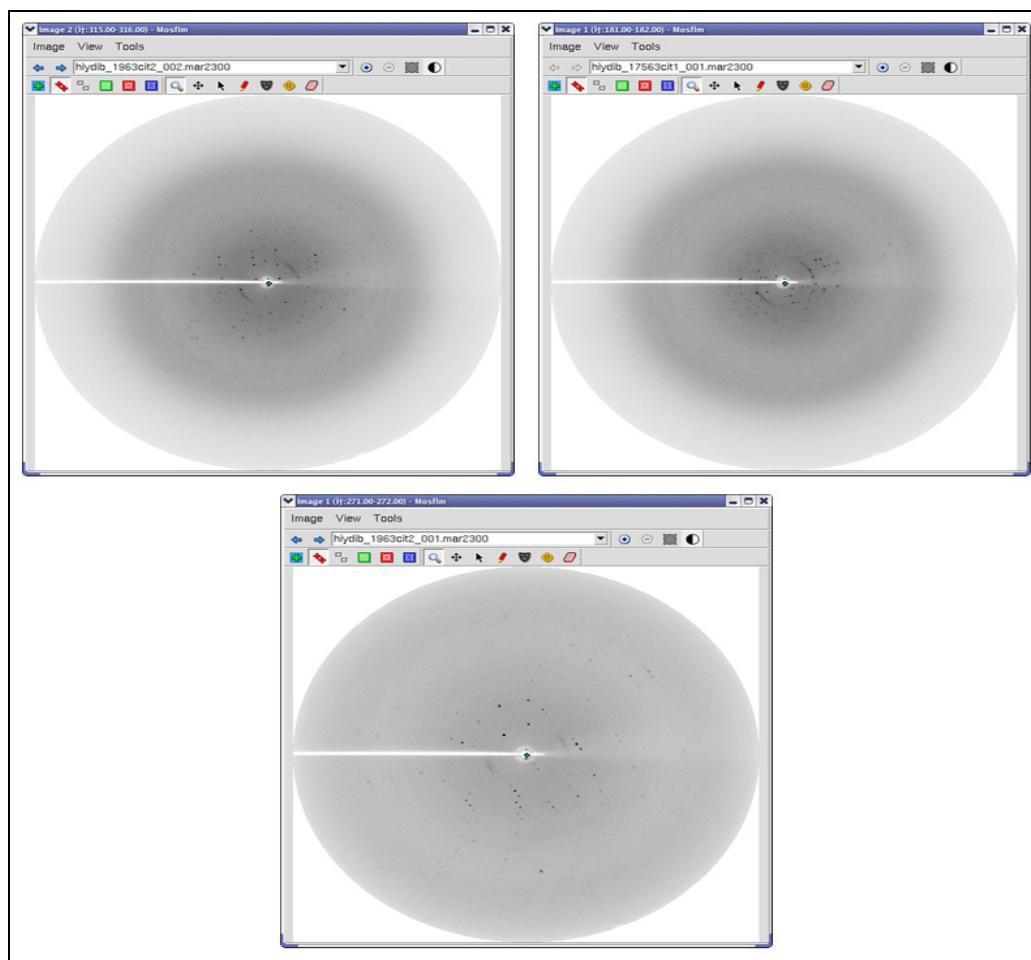


Figure 12.11 Diffraction pattern for *H/SdhL* crystals from C.No 4, 6 & 9

## 12.3 Conclusions and future work

The co-crystallisation and diffraction studies were tried until it became clear that no improvements were going to be forthcoming. The absence of an over expressing PTB361 clone of *H/SdhL* stopped the possibility of purifying the native protein, which may have resulted in better diffraction of the crystals and ultimately a crystal structure in the presence of Mannose 6-phosphate. It is well known that the histidine tag can be detrimental to the success of crystallization

trials. Future work should focus on obtaining diffraction quality crystals which could diffract to  $\sim 2.0\text{\AA}$  which has been achieved previously (Kirsty Stewart, PhD Thesis 2005 University of Glasgow). The soaking of crystals with mannose-6-phosphate would be the simplest approach for obtaining a complex crystal structure. Another route and potentially time consuming would be to use NMR to solve the solution structure, however this would have been beyond the scope of this thesis.

## 13. References

1. de Souza, N. (2007) From structure to function, *Nature Methods* 4, 771-771.
2. Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K., and Raushel, F. M. (2007) Structure-based activity prediction for an enzyme of unknown function, *Nature* 448, 775-U772.
3. Gerlt, J. A. (2007) A protein structure (or function?) initiative, *Structure* 15, 1353-1356.
4. Devos, D., and Valencia, A. (2000) Practical limits of function prediction, *Proteins-Structure Function and Genetics* 41, 98-107.
5. Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective, *Journal of Molecular Biology* 307, 1113-1143.
6. Campbell, S. J., Gold, N. D., Jackson, R. M., and Westhead, D. R. (2003) Ligand binding: functional site location, similarity and docking, *Current Opinion in Structural Biology* 13, 389-395.
7. Laskowski, R. A., Luscombe, N. M., Swindells, M. B., and Thornton, J. M. (1996) Protein clefts in molecular recognition and function, *Protein Science* 5, 2438-2452.
8. Sotriffer, C., and Klebe, G. (2002) Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design, *Farmaco* 57, 243-251.
9. Marsden, P. M., Puvanendrapillai, D., Mitchell, J. B. O., and Glen, R. C. (2004) Predicting protein-ligand binding affinities: a low scoring game?, *Organic & Biomolecular Chemistry* 2, 3267-3273.
10. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and clustal X version 2.0, *Bioinformatics* 23, 2947-2948.
11. Favia, A. D., Nobeli, I., Glaser, F., and Thornton, J. M. (2008) Molecular docking for substrate identification: The short-chain dehydrogenases/reductases, *Journal of Molecular Biology* 375, 855-874.
12. Seffernick, J. L., de Souza, M. L., Sadowsky, M. J., and Wackett, L. P. (2001) Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different, *Journal of Bacteriology* 183, 2405-2410.
13. Glasner, M. E., Fayazmanesh, N., Chiang, R., Jacobson, M. P., Gerlt, J. A., and Babbitt, P. C. (2006) Evolution of structure and function in the o-succinylbenzoate synthase family, *Faseb J.* 20, A905-A905.
14. Nguyen, T. T., Brown, S., Fedorov, A. A., Fedorov, E. V., Babbitt, P. C., Almo, S. C., and Raushel, F. M. (2008) At the Periphery of the Amidohydrolase Superfamily: Bh0493 from *Bacillus halodurans* Catalyzes the Isomerization of d-Galacturonate to d-Tagaturonate, *Biochemistry* 47, 1194-1206.
15. Gerlt, J. A., and Babbitt, P. C. (2000) Can sequence determine function?, *Genome Biology* 1.
16. Anfinsen, C. B. (1973) Principles That Govern Folding of Protein Chains, *Science* 181, 223-230.
17. Moulton, J., and Melamud, E. (2000) From fold to function, *Current Opinion in Structural Biology* 10, 384-389.

18. Chothia, C., and Lesk, A. M. (1986) The Relation between the Divergence of Sequence and Structure in Proteins, *Embo Journal* 5, 823-826.
19. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Research* 28, 235-242.
20. Berman, H. M., and Westbrook, J. D. (2004) The impact of structural genomics on the protein data bank, *American journal of pharmacogenomics : genomics-related research in drug development and clinical practice* 4, 247-252.
21. Laskowski, R. A., Watson, J. D., and Thornton, J. M. (2005) ProFunc: a server for predicting protein function from 3D structure, *Nucleic Acids Research* 33, W89-W93.
22. Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA, *Proteins Suppl* 3, 171-176.
23. Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J., and Orengo, C. (2003) Recognizing the fold of a protein structure, *Bioinformatics* 19, 1748-1759.
24. Krishna, S. S., Weekes, D., Bakolitsa, C., Elsliger, M.-A., Wilson, I. A., Godzik, A., and Wooley, J. (2010) TOPSAN: use of a collaborative environment for annotating, analyzing and disseminating data on JCSG and PSI structures, *Acta Crystallographica Section F-Structural Biology and Crystallization Communications* 66, 1143-1147.
25. Devos, D., and Valencia, A. (2001) Intrinsic errors in genome annotation, *Trends in Genetics* 17, 429-431.
26. Whisstock, J. C., and Lesk, A. M. (2003) Prediction of protein function from protein sequence and structure, *Quarterly Reviews of Biophysics* 36, 307-340.
27. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D. W., Sali, A., Studier, F. W., and Swaminathan, S. (1999) Structural genomics: beyond the Human Genome Project, *Nature Genetics* 23, 151-157.
28. Brenner, S. E. (2001) A tour of structural genomics, *Nature Reviews Genetics* 2, 801-809.
29. Stevens, R. C., Yokoyama, S., and Wilson, I. A. (2001) Global efforts in structural genomics, *Science* 294, 89-92.
30. Chance, M. R., Bresnick, A. R., Burley, S. K., Jiang, J. S., Lima, C. D., Sali, A., Almo, S. C., Bonanno, J. B., Buglino, J. A., Boulton, S., Chen, H., Eswar, N., He, G. S., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M., and Wang, L. K. (2002) Structural genomics: A pipeline for providing structures for the biologist, *Protein Science* 11, 723-738.
31. Watson, J. D., Laskowski, R. A., and Thornton, J. M. (2005) Predicting protein function from sequence and structural data, *Current Opinion in Structural Biology* 15, 275-284.
32. Brenner, S. E. (1999) Errors in genome annotation, *Trends in Genetics* 15, 132-133.
33. Chang, C., Ekins, S., Bahadduri, P., and Swaan, P. W. (2006) Pharmacophore-based discovery of ligands for drug transporters, *Advanced Drug Delivery Reviews* 58, 1431-1450.
34. Friedberg, I., Jambon, M., and Godzik, A. (2006) New avenues in protein function prediction, *Protein Science* 15, 1527-1529.

35. Böhm, H. J. (2005) Prediction of Non-bonded Interactions in Drug Design, In *Protein-Ligand Interactions*, pp 3-20, Wiley-VCH Verlag GmbH & Co. KGaA.
36. Southall, N. T., Dill, K. A., and Haymet, A. D. J. (2002) A view of the hydrophobic effect, *Journal of Physical Chemistry B* 106, 521-533.
37. Snyder, P. W., Mecinovic, J., Moustakas, D. T., Thomas, S. W., III, Harder, M., Mack, E. T., Lockett, M. R., Heroux, A., Sherman, W., and Whitesides, G. M. (2011) Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase, *Proceedings of the National Academy of Sciences of the United States of America* 108, 17889-17894.
38. Brenk, R., and Klebe, G. (2006) "Hot Spot" Analysis of Protein-binding Sites as a Prerequisite for Structure-based Virtual Screening and Lead Optimization, In *Pharmacophores and Pharmacophore Searches*, pp 171-192, Wiley-VCH Verlag GmbH & Co. KGaA.
39. A Ben-Naim Lapanje, S. (1980) Hydrophobic interactions: By A Ben-Naim. pp 311. Plenum Press: New York and London. 1980. \$32.50 ISBN 0-306-40222-X, *Biochemical Education* 8, 124-124.
40. Williams, M. A., and Ladbury, J. E. (2005) Hydrogen Bonds in Protein-Ligand Complexes, In *Protein-Ligand Interactions*, pp 137-161, Wiley-VCH Verlag GmbH & Co. KGaA.
41. Höfliger, M. M., and Beck-Sickinger, A. G. (2005) Receptor-Ligand Interaction, In *Protein-Ligand Interactions*, pp 107-135, Wiley-VCH Verlag GmbH & Co. KGaA.
42. Raffa, R. B. (2005) Experimental Approaches to Determine the Thermodynamics of Protein-Ligand Interactions, In *Protein-Ligand Interactions*, pp 51-71, Wiley-VCH Verlag GmbH & Co. KGaA.
43. Kirchmair, J., Spitzer, G. M., and Liedl, K. R. (2010) Consideration of Water and Solvation Effects in Virtual Screening, In *Virtual Screening*, pp 263-289, Wiley-VCH Verlag GmbH & Co. KGaA.
44. Lengauer, T., and Rarey, M. (1996) Computational methods for biomolecular docking, *Current Opinion in Structural Biology* 6, 402-406.
45. Joy, S., Nair, P. S., Hariharan, R., and Pillai, M. R. (2006) Detailed comparison of the protein-ligand docking efficiencies of GOLD, a commercial package and ArgusLab, a licensable freeware, *In silico biology* 6, 601-605.
46. Hu, Z. J., Wang, S. M., and Southerland, W. M. (2004) Dock-odysseys. II. The evaluation of autodock program and its comparison with other docking programs, *Abstracts of Papers of the American Chemical Society* 228, U361-U361.
47. Lemmen, C., Hindle, S. A., Gastreich, M., Dramburg, I., and Claussen, H. (2004) FlexX-docking: Past, present and planned technological advancements, *Abstracts of Papers of the American Chemical Society* 228, U507-U507.
48. Guener, O. F. (2000) *Pharmacophore: Perception, Development, and Use in Drug Design*. [In: *IUL Biotechnol. Ser.*, 2000; 2], International University Line.
49. Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004) Docking and scoring in virtual screening for drug discovery: Methods and applications, *Nature Reviews Drug Discovery* 3, 935-949.
50. Song, L., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Brown, S., Imker, H. J., Babbitt, P. C., Almo, S. C., Jacobson, M. P., and Gerlt, J. A. (2007) Prediction and assignment of function for a

- divergent N-succinyl amino acid racemase, *Nature Chemical Biology* 3, 486-491.
51. Rantanen, V. V., Denessiouk, K. A., Gyllenberg, M., Koski, T., and Johnson, M. S. (2001) A fragment library based on gaussian mixtures predicting favorable molecular interactions, *Journal of Molecular Biology* 313, 197-214.
  52. Hindle, S. A., Rarey, M., Buning, C., and Lengauer, T. (2002) Flexible docking under pharmacophore type constraints, *Journal of Computer-Aided Molecular Design* 16, 129-149.
  53. Burger, A. (1970) *History and economics of medicinal chemistry*.
  54. Gund, K., P. (1977) Three dimensional pharmacophoric pattern searching In: *Progress in Molecular and Subcellular Biology*, herausgegeben von F. E. Hahn, T. T. Puck, G. F. Springer und K. Wallenfels. Seiten, zahl-reiche Abb., Springer-Verlag, Heidelberg, New York 5, 117-143.
  55. Wermuth, G., Ganellin, C. R., Lindberg, P., and Mitscher, L. A. (1998) Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998), *Pure and Applied Chemistry* 70, 1129-1143.
  56. Pickett, S. (2005) The Biophore Concept, In *Protein-Ligand Interactions*, pp 73-105, Wiley-VCH Verlag GmbH & Co. KGaA.
  57. Mony, L., Triballeau, N., Paoletti, P., Acher, F. C., and Bertrand, H.-O. (2010) Identification of a novel NR2B-selective NMDA receptor antagonist using a virtual screening approach, *Bioorganic & Medicinal Chemistry Letters* 20, 5552-5558.
  58. Chang, C., and Ekins, S. (2006) Pharmacophores for Human ADME/Tox-Related Proteins, In *Pharmacophores and Pharmacophore Searches*, pp 299-324, Wiley-VCH Verlag GmbH & Co. KGaA.
  59. Langer, T., Eder, M., Hoffmann, R. D., Chiba, P., and Ecker, G. F. (2004) Lead Identification for Modulators of Multidrug Resistance based on in silico Screening with a Pharmacophoric Feature Model, *Archiv der Pharmazie* 337, 317-327.
  60. Langer, T., and Hoffman, R. D. (2006) Pharmacophores and Pharmacophore Searches, *Methods and Principles in Medicinal Chemistry*, WILEY-VCH 32.
  61. Al-Nadaf, A., Sheikha, G. A., and Taha, M. O. (2010) Elaborate ligand-based pharmacophore exploration and QSAR analysis guide the synthesis of novel pyridinium-based potent  $\beta$ -secretase inhibitory leads, *Bioorganic & Medicinal Chemistry* 18, 3088-3115.
  62. Steindl, T., and Langer, T. (2004) Influenza virus neuraminidase inhibitors: generation and comparison of structure-based and common feature pharmacophore hypotheses and their application in virtual screening, *J Chem Inf Comput Sci* 44, 1849-1856.
  63. Carlson, H. A., Masukawa, K. M., Rubins, K., Bushman, F. D., Jorgensen, W. L., Lins, R. D., Briggs, J. M., and McCammon, J. A. (2000) Developing a Dynamic Pharmacophore Model for HIV-1 Integrase, *Journal of Medicinal Chemistry* 43, 2100-2114.
  64. Wermuth, C. G. (2006) Pharmacophores: Historical Perspective and Viewpoint from a Medicinal Chemist, In *Pharmacophores and Pharmacophore Searches*, pp 1-13, Wiley-VCH Verlag GmbH & Co. KGaA.
  65. Wolber, G., Kirchmair, J., and Langer, T. (2005) Structure -Based 3D Pharmacophores: An Alternative to Docking?, 7<sup>th</sup> international Conference on chemical structures. June 5 - 9, 2005 Noordwijkerhout, The Netherlands, [www.int-conf-chem-structures.org](http://www.int-conf-chem-structures.org).



66. Wolber, G., and Langer, T. (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters, *J Chem Inf Model* 45, 160-169.
67. Accelrys. DS ViewerPro. Accelrys, Inc., San Diego, CA.
68. Accelrys, C. [http://www.accelrys.com/catalyst/cat\\_info.html](http://www.accelrys.com/catalyst/cat_info.html).
69. Cerius<sup>2</sup>, and Catalyst. are software packages available from molecular simulations Inc., 9685 Scranton Rd., San Diego, CA 92121-3752.
70. Böhm, H.-J. (1992) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads, *Journal of Computer-Aided Molecular Design* 6, 593-606.
71. Greenidge, P. A., Merette, S. A., Beck, R., Dodson, G., Goodwin, C. A., Scully, M. F., Spencer, J., Weiser, J., and Deadman, J. J. (2003) Generation of ligand conformations in continuum solvent consistent with protein active site topology: application to thrombin, *J Med Chem* 46, 1293-1305.
72. Gillner, M., and Greenidge, P. (1999) Pharmacophores including multiple excluded volumes derived from X-ray crystallographic target structures to be used in 3D-database searching, *Abstracts of Papers of the American Chemical Society* 217, U566-U566.
73. Ekins, S., and Swaan, P. W. (2004) Development of computational models for enzymes, transporters, channels, and receptors relevant to ADME/Tox, *Reviews in Computational Chemistry, Vol 20* 20, 333-415.
74. Markt, P., Schuster, D., and Langer, T. (2010) Pharmacophore Models for Virtual Screening, In *Virtual Screening*, pp 115-152, Wiley-VCH Verlag GmbH & Co. KGaA.
75. Triballeau, N., Bertrand, H.-O., and Acher, F. (2006) Are You Sure You Have a Good Model?, In *Pharmacophores and Pharmacophore Searches*, pp 325-364, Wiley-VCH Verlag GmbH & Co. KGaA.
76. Wolber, G., and Kosara, R. (2006) Pharmacophores from Macromolecular Complexes with LigandScout, In *Pharmacophores and Pharmacophore Searches*, pp 131-150, Wiley-VCH Verlag GmbH & Co. KGaA.
77. Drenth, J. (2007) *Principles of Protein X-Ray Crystallography, Third Edition*.
78. Dhaliwal, B., Nichols, C. E., Ren, J. S., Lockyer, M., Charles, I., Hawkins, A. R., and Stammers, D. K. (2004) Crystallographic studies of shikimate binding and induced conformational changes in Mycobacterium tuberculosis shikimate kinase, *Febs Letters* 574, 49-54.
79. Tanizawa, K., Matsunami, H., and Yamaguchi, H. (2000) *Mechanism of topa quinone biogenesis in copper amine oxidase studied by site-directed mutagenesis and x-ray crystallography*.
80. Gerlt, J. A., Kozarich, J. W., Kenyon, G. L., Neidhart, D. J., and Petsko, G. A. (1992) *Mechanism of the reaction of catalyzed by mandelate racemase: Lessons from protein engineering, chemical modification, and X-ray crystallography*.
81. Bartlett, P. A., Sampson, N. S., Reich, S. H., Drewry, D. H., and Lamden, L. A. (1990) *Interplay among enzyme mechanism protein structure and the design of serine protease inhibitors*.
82. Dodson, E., and Dodson, G. (2009) Movements at the Hemoglobin A-Hemes and their Role in Ligand Binding, Analyzed by X-Ray Crystallography, *Biopolymers* 91, 1056-1063.
83. Viegas, A., Bras, N. F., Cerqueira, N. M. F. S. A., Fernandes, P. A., Prates, J. A. M., Fontes, C. M. G. A., Bruix, M., Romao, M. J., Carvalho, A. L., Ramos, M. J., Macedo, A. L., and Cabrita, E. J. (2008) Molecular



- determinants of ligand specificity in family 11 carbohydrate binding modules - an NMR, X-ray crystallography and computational chemistry approach, *Febs Journal* 275, 2524-2535.
84. Bjerrum, E. J., and Biggin, P. C. (2008) Rigid body essential X-ray crystallography: Distinguishing the bend and twist of glutamate receptor ligand binding domains, *Proteins-Structure Function and Bioinformatics* 72, 434-446.
85. Goldstein, S. M., Bordner, J., Hoth, L. R., and Geoghegan, K. F. (2001) Chemical and biochemical issues related to X-ray crystallography of the ligand-binding domain of estrogen receptor alpha, *Bioconjugate Chemistry* 12, 406-413.
86. Gouaux, E., and Armstrong, N. (2000) Action of agonists and antagonists on the AMPA GluR2 receptor ligand binding core defined by x-ray crystallography, *Abstracts of Papers of the American Chemical Society* 219, U296-U296.
87. Cameron, A. D., Smerdon, S. J., Wilkinson, A. J., Habash, J., Helliwell, J. R., Li, T. S., and Olson, J. S. (1993) Distal Pocket Polarity in Ligand-Binding to Myoglobin - Deoxy and Carbonmonoxy Forms of a Threonine(68)(E11) Mutant Investigated by X-Ray Crystallography and Infrared-Spectroscopy, *Biochemistry* 32, 13061-13070.
88. Edmundson, A. B., Herron, J. N., Ely, K. R., Harris, D. L., Voss, E. W. J., Tribbick, G., and Geysen, H. M. (1990) *Complexes of peptide nucleotides and flouresein with immunoglobulin fragments effects of solvent on crystal structure and ligand binding.*
89. Frase, H., Smith, C. A., Toth, M., Champion, M. M., Mobashery, S., and Vakulenko, S. B. (2011) Identification of Products of Inhibition of GES-2 beta-Lactamase by Tazobactam by X-ray Crystallography and Spectrometry, *Journal of Biological Chemistry* 286, 14396-14409.
90. Cuerrier, D., Moldoveanu, T., Inoue, J., Davies, P. L., and Campbell, R. L. (2006) Calpain inhibition by alpha-ketoamide and cyclic hemiacetal inhibitors revealed by X-ray crystallography, *Biochemistry* 45, 7446-7452.
91. Veach, D. R., Swendeman, S., Nagar, B., Wisniewski, D., Strife, A., Lambek, C. L., Liu, C.-Y., Lee, W. W., Bommann, W. G., Kuriyan, J., Bertino, J., and Clarkson, B. (2002) Towards picomolar inhibition of Bcr-Abl: Synthesis and evaluation of a focused library of pyrido-(2,3-d)-pyrimidine tyrosine kinase inhibitors guided by x-ray crystallography and molecular modeling, *Proceedings of the American Association for Cancer Research Annual Meeting* 43, 847.
92. Steiner, R. A., Kooter, I. M., and Dijkstra, B. W. (2002) Functional analysis of the copper-dependent quercetin 2,3-dioxygenase. 1. Ligand-induced coordination changes probed by X-ray crystallography: Inhibition, ordering effect, and mechanistic insights, *Biochemistry* 41, 7955-7962.
93. Huang, K. F., Chiou, S. H., Ko, T. P., and Wang, A. H. J. (2002) Determinants of the inhibition of a Taiwan habu venom metalloproteinase by its endogenous inhibitors revealed by X-ray crystallography and synthetic inhibitor analogues, *European Journal of Biochemistry* 269, 3047-3056.
94. Benach, J., Atrian, S., Gonzalez-Duarte, R., and Ladenstein, R. (1999) The catalytic reaction and inhibition mechanism of Drosophila alcohol dehydrogenase: Observation of an enzyme-bound NAD-ketone adduct at 1.4 angstrom resolution by X-ray crystallography, *Journal of Molecular Biology* 289, 335-355.

95. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics, *Acta Crystallographica Section D-Biological Crystallography* 60, 2126-2132.
96. Weiss, M. (2012) Crystals, X-rays and Proteins: Comprehensive Protein Crystallography. By Dennis Sherwood and Jon Cooper. Oxford University Press, 2010. Pp. 626. Price USD 98.50. ISBN 978-01995-5904-6, *Acta crystallographica. Section D, Biological crystallography* 68, 93-94.
97. Cooper, A. (2004) *Biophysical chemistry*, Royal Society of Chemistry.
98. Fischer, J. J., and Jardetzky, O. (1965) Nuclear Magnetic Relaxation Study of Intermolecular Complexes. The Mechanism of Penicillin Binding to Serum Albumin, *J Am Chem Soc* 87, 3237-3244.
99. Thewes, T., Constantine, K., Byeon, I. J. L., and Llinas, M. (1990) Ligand Interactions with the Kringle-5 Domain of Plasminogen - a Study by H-1-Nmr Spectroscopy, *Journal of Biological Chemistry* 265, 3906-3915.
100. Lepre, C. A. (2001) Library design for NMR-based screening, *Drug Discovery Today* 6, 133-140.
101. Hajduk, P. J., Gomtsyan, A., Didomenico, S., Cowart, M., Bayburt, E. K., Solomon, L., Severin, J., Smith, R., Walter, K., Holzman, T. F., Stewart, A., McGaraughty, S., Jarvis, M. F., Kowaluk, E. A., and Fesik, S. W. (2000) Design of Adenosine Kinase Inhibitors from the NMR-Based Screening of Fragments, *Journal of Medicinal Chemistry* 43, 4781-4786.
102. Doucleff, M., HatcherSkeers, M., and Crane, N. J. (2011) *Pocket Guide to Biomolecular NMR*.
103. Akke, M. (2012) Conformational dynamics and thermodynamics of protein-ligand binding studied by NMR relaxation, *Biochemical Society transactions* 40, 419-423.
104. Parsons, L., Bonander, N., Eisenstein, E., Gilson, M., Kairys, V., and Orban, J. (2003) Solution structure and functional ligand screening of HI0719, a highly conserved protein from bacteria to humans in the YjgF/YER057c/UK114 family, *Biochemistry* 42, 80-89.
105. Paramanik, V., and Thakur, M. K. (2011) NMR analysis reveals 17beta-estradiol induced conformational change in ERbeta ligand binding domain expressed in E. coli, *Mol Biol Rep* 38, 4657-4661.
106. Macarrón, R., and Hertzberg, R. P. (2009) Design and Implementation of High-Throughput Screening Assays, (Janzen, W. P., and Bernasconi, P., Eds.), pp 1-32, Humana Press.
107. Williams, K. P., and Scott, J. E. (2009) Enzyme Assay Design for High-Throughput Screening, (Janzen, W. P., and Bernasconi, P., Eds.), pp 107-126, Humana Press.
108. Sonawane, N. D., Zhao, D., Zegarra-Moran, O., Galiotta, L. J. V., and Verkman, A. S. (2008) Nanomolar CFTR inhibition by pore-occluding divalent polyethylene glycol-malonic acid hydrazides, *Chemistry & Biology* 15, 718-728.
109. More, S. S., and Vince, R. (2009) Inhibition of Glyoxalase I: The First Low-Nanomolar Tight-Binding Inhibitors, *Journal of Medicinal Chemistry* 52, 4650-4656.
110. Ritschel, T., Kohler, P. C., Neudert, G., Heine, A., Diederich, F., and Klebe, G. (2009) How to Replace the Residual Solvation Shell of Polar Active Site Residues to Achieve Nanomolar Inhibition of tRNA-Guanine Transglycosylase, *Chemmedchem* 4, 2012-2023.
111. Paskaleva, E. E., Xue, J., Lee, D. Y. W., Shekhtman, A., and Canki, M. (2010) Palmitic Acid Analogs Exhibit Nanomolar Binding Affinity for the

- HIV-1 CD4 Receptor and Nanomolar Inhibition of gp120-to-CD4 Fusion, *Plos One* 5.
112. Nidetzky, B., Mayr, P., Hadwiger, P., and Stutz, A. E. (1999) Binding energy and specificity in the catalytic mechanism of yeast aldose reductases, *Biochemical Journal* 344, 101-107.
  113. Zhang, X., Chen, L., Fei, X.-C., Ma, Y.-S., and Gao, H.-W. (2009) Binding of PFOS to serum albumin and DNA: insight into the molecular toxicity of perfluorochemicals, *Bmc Molecular Biology* 10.
  114. Czodrowski, P., Sottriffer, C. A., and Klebe, G. (2007) Protonation Changes upon Ligand Binding to Trypsin and Thrombin: Structural Interpretation Based on pKa Calculations and ITC Experiments, *Journal of Molecular Biology* 367, 1347-1356.
  115. Meulen, K. A. V., and Butcher, S. E. (2012) Characterization of the kinetic and thermodynamic landscape of RNA folding using a novel application of isothermal titration calorimetry, *Nucleic Acids Research* 40, 2140-2151.
  116. Cooper, A. (1999) Thermodynamic analysis of biomolecular interactions, *Current Opinion in Chemical Biology* 3, 557-563.
  117. Waterhous, D. V., Brouillette, C. G., Fish, F., and Muccio, D. D. (1989) Spectroscopic and DSC studies provide evidence for stable intermediates in the thermal unfolding and refolding pathway of serum retinol binding protein, *Biophys. J.* 55, 414.
  118. Domach, M. M., and Sagar, S. L. (1995) Employing Dsc for Elucidating Metal-Protein Binding Modalities, *Abstracts of Papers of the American Chemical Society* 209, 64-Btec.
  119. Straume, M., and Freire, E. (1992) 2-Dimensional Differential Scanning Calorimetry (2d Dsc) - Global Linkage Analysis Allows Simultaneous Resolution of Ligand-Binding Interactions and Intrinsic Protein Structural Energetics, *Faseb J.* 6, A168-A168.
  120. Kelly, S. M., Jess, T. J., and Price, N. C. (2005) How to study proteins by circular dichroism, *Biochimica Et Biophysica Acta-Proteins and Proteomics* 1751, 119-139.
  121. Lin, S., Xu, M., and Yuan, G. (2012) Study of STAT3 G-quadruplex folding patterns by CD spectroscopy and molecular modeling, *Chin. Chem. Lett.* 23, 329-331.
  122. Amiri, R., Bordbar, A. K., Garcia-Mayoral, M., Khosropour, A. R., Mohammadpoor-Baltork, I., Menendez, M., and Laurents, D. V. (2012) Interactions of gemini surfactants with two model proteins: NMR, CD, and fluorescence spectroscopies, *J. Colloid Interface Sci.* 369, 245-255.
  123. Jain, N., Meneni, S. R., and Cho, B. (2007) Sequence effects on the NarI-derived frameshift mutagenesis by 19F NMR, DSC, and CD spectroscopy, *Abstracts of Papers of the American Chemical Society* 234.
  124. Studier, F. W., and Moffatt, B. A. (1986) Use of Bacteriophage-T7 Rna-Polymerase to Direct Selective High-Level Expression of Cloned Genes, *Journal of Molecular Biology* 189, 113-130.
  125. Meselson, M., and Yuan, R. (1968) DNA restriction enzyme from E. coli, *Nature* 217, 1110-1114.
  126. Novagen. (2006) Novagen, pETSystemManual
  127. Gallagher, S. R. (2011) Quantitation of DNA and RNA with absorption and fluorescence spectroscopy, *Current protocols in neuroscience / editorial board, Jacqueline N. Crawley ... [et al.] Appendix 1*, 1K.
  128. Inoue, H., Nojima, H., and Okayama, H. (1990) High efficiency transformation of Escherichia coli with plasmids, *Gene* 96, 23-28.

129. Gaberc-Porekar, V., and Menart, V. (2001) Perspectives of immobilized-metal affinity chromatography, *Journal of Biochemical and Biophysical Methods* 49, 335-360.
130. Sulkowski, E. (1996) Immobilized metal-ion affinity chromatography: Imidazole proton pump and chromatographic sequelae .1. Proton pump, *J. Mol. Recognit.* 9, 389-393.
131. Gill, S. C., and von Hippel, P. H. (1989) Calculation of protein extinction coefficients from amino acid sequence data, *Anal Biochem* 182, 319-326.
132. Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. T. (1960) Structure of Haemoglobin - 3-Dimensional Fourier Synthesis at 5.5-Å Resolution, Obtained by X-Ray Analysis, *Nature* 185, 416-422.
133. Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., and Shore, V. C. (1960) Structure of Myoglobin - 3-Dimensional Fourier Synthesis at 2 Å Resolution, *Nature* 185, 422-427.
134. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest, *Nucleic Acids Research* 36, D344-D350.
135. Tarle, I., Borhani, D. W., Wilson, D. K., Quijcho, F. A., and Petrash, J. M. (1993) Probing the active site of human aldose reductase. Site-directed mutagenesis of Asp-43, Tyr-48, Lys-77, and His-110, *The Journal of biological chemistry* 268, 25687-25693.
136. Kavanagh, K. L., Klimacek, M., Nidetzky, B., and Wilson, D. K. (2002) The structure of apo and holo forms of xylose reductase, a dimeric aldo-keto reductase from *Candida tenuis*, *Biochemistry* 41, 8785-8795.
137. Kador, P. F., Robison, W. G., Jr., and Kinoshita, J. H. (1985) The pharmacology of aldose reductase inhibitors, *Annu Rev Pharmacol Toxicol* 25, 691-714.
138. Neuhauser, W., Haltrich, D., Kulbe, K. D., and Nidetzky, B. (1997) NAD(P)H-dependent aldose reductase from the xylose-assimilating yeast *Candida tenuis* - Isolation, characterization and biochemical properties of the enzyme, *Biochemical Journal* 326, 683-692.
139. Kratzer, R., Wilson, D. K., and Nidetzky, B. (2006) Catalytic mechanism and substrate selectivity of aldo-keto reductases: Insights from structure-function studies of *Candida tenuis* xylose reductase, *lubmb Life* 58, 499-507.
140. Harrison, D. H., Bohren, K. M., Ringe, D., Petsko, G. A., and Gabbay, K. H. (1994) ANION-BINDING SITE IN HUMAN ALDOSE REDUCTASE - MECHANISTIC IMPLICATIONS FOR THE BINDING OF CITRATE, CACODYLATE, AND GLUCOSE-6-PHOSPHATE, *Biochemistry* 33, 2011-2020.
141. Mayr, P., and Nidetzky, B. (2002) Catalytic reaction profile for NADH-dependent reduction of aromatic aldehydes by xylose reductase from *Candida tenuis*, *Biochemical Journal* 366, 889-899.
142. Krell, T., Coggins, J. R., and Laphorn, A. J. (1998) The three-dimensional structure of shikimate kinase, *Journal of Molecular Biology* 278, 983-997.
143. Gu, Y. J., Reshetnikova, L., Li, Y., Wu, Y., Yan, H. G., Singh, S., and Ji, X. H. (2002) Crystal structure of shikimate kinase from *Mycobacterium tuberculosis* reveals the dynamic role of the LID domain in catalysis, *Journal of Molecular Biology* 319, 779-789.
144. WHO, and Report, (Eds.) (1999) *World Health Organization, Report on Infectious Diseases: Removing Obstacles to Healthy Development.*, Atar, Switzerland.

145. Burman, J. D., Stevenson, C. E. M., Sawers, R. G., and Lawson, D. M. (2007) The crystal structure of *Escherichia coli* TdcF, a member of the highly conserved YjgF/YER057c/UK114 family, *Bmc Structural Biology* 7.
146. Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, P., Ulrich, E. L., Markley, J. L., Ionides, J., and Laue, E. D. (2005) The CCPN data model for NMR spectroscopy: Development of a software pipeline, *Proteins-Structure Function and Bioinformatics* 59, 687-696.
147. Masterson, L. R., Mascioni, A., Traaseth, N. J., Taylor, S. S., and Veglia, G. (2008) Allosteric cooperativity in protein kinase A, *Proceedings of the National Academy of Sciences of the United States of America* 105, 506-511.
148. Muller, O. H., and Baumberger, J. P. (1939) The keto-enol tautomerism of pyruvate ion studied polarographically, *Journal of the American Chemical Society* 61, 590-596.
149. Burman, J. D., Stevenson, C. E. M., Hauton, K. A., Sawers, G., and Lawson, D. M. (2003) Crystallization and preliminary X-ray analysis of the *E.-coli* hypothetical protein TdcF, *Acta Crystallographica Section D-Biological Crystallography* 59, 1076-1078.
150. Zhang, X.-X., and Rainey, P. B. (2007) Genetic analysis of the histidine utilization (hut) genes in *Pseudomonas fluorescens* SBW25, *Genetics* 176, 2165-2176.
151. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. (2000) The EcoCyc and MetaCyc databases, *Nucleic Acids Research* 28, 56-59.
152. Donovan, R. S., Robinson, C. W., and Glick, B. R. (1996) Review: Optimizing inducer and culture conditions for expression of foreign proteins under the control of the lac promoter, *Journal of Industrial Microbiology* 16, 145-154.
153. Mitraki, A., and King, J. (1989) Protein Folding Intermediates and Inclusion Body Formation, *Bio-Technology* 7, 690-697.
154. Liu, Y. (2009) Structural and biochemical analysis of HutD from *Pseudomonas fluorescens* SBW25 : a thesis submitted in fulfilment of the requirements for the degree of Master of Science in Molecular Biosciences at Massey University, Auckland, New Zealand.
155. Hayaishi, O., Tabor, H., and Hayaishi, T. (1957) N-Formimino-L-Aspartic Acid as an Intermediate in the Enzymatic Conversion of Imidazoleacetic Acid to Formylaspartic Acid, *Journal of Biological Chemistry* 227, 161-180.
156. (1958) Biochemical preparations. Vol. 5. Edited by the 20-member editorial board with David Shemin as Editor-in-Chief. John Wiley and Sons, Inc., N. Y., 1957. ix + 115 pp. 15 × 23 cm. Price \$4.75, *Journal of the American Pharmaceutical Association* 47, 762-762.
157. Berezhkovskii, A. M., Szabo, A., and Weiss, G. H. (2000) Theory of the Fluorescence of Single Molecules Undergoing Multistate Conformational Dynamics†, *The Journal of Physical Chemistry B* 104, 3776-3780.
158. Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering, *Nucleic Acids Res* 16, 10881-10890.
159. Harel, M., Kleywegt, G. J., Ravelli, R. B. G., Silman, I., and Sussman, J. L. (1995) Crystal structure of an acetylcholinesterase-fasciculin complex: interaction of a three-fingered toxin from snake venom with its target, *Structure* 3, 1355-1366.
160. Kleywegt, G. J. (1999) Recognition of spatial motifs in protein structures, *Journal of Molecular Biology* 285, 1887-1897.

161. Novotny, M., and Kleywegt, G. J. (2005) A Survey of Left-handed Helices in Protein Structures, *Journal of Molecular Biology* 347, 231-241.
162. Schineller, J. B. (1999) An Introduction to Enzyme and Coenzyme Chemistry (Bugg, Tim), *Journal of Chemical Education* 76, 1070.
163. Abramowitz, N., Schechter, I., and Berger, A. (1967) On the size of the active site in proteases II. Carboxypeptidase-A, *Biochemical and Biophysical Research Communications* 29, 862-867.
164. Verardo, G., Geatti, P., and Strazzolini, P. (2007) Rapid and efficient microwave-assisted synthesis of N-carbamoyl-L-amino acids, *Synthetic Communications* 37, 1833-1844.
165. Rossmore, H. W., and Sondossi, M. (1988) Applications and Mode of Action of Formaldehyde Condensate Biocides, In *Advances in Applied Microbiology* (Allen, I. L., Ed.), pp 223-277, Academic Press.
166. Choquet, C. G., Richards, J. C., Patel, G. B., and Sprott, G. D. (1994) Purine and pyrimidine biosynthesis in methanogenic bacteria, *Archives of Microbiology* 161, 471-480.
167. Cole, J. C., Korb, O., Olsson, T. S. G., and Liebeschuetz, J. (2010) The Basis for Target-Based Virtual Screening: Protein Structures, In *Virtual Screening*, pp 87-114, Wiley-VCH Verlag GmbH & Co. KGaA.
168. Davis, A. M., Teague, S. J., and Kleywegt, G. J. (2003) Application and Limitations of X-ray Crystallographic Data in Structure-Based Ligand and Drug Design, *Angewandte Chemie International Edition* 42, 2718-2736.
169. Hooft, R. W. W., Vriend, G., Sander, C., and Abola, E. E. (1996) Errors in protein structures, *Nature* 381, 272-272.
170. Nissink, J. W. M., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C., and Taylor, R. (2002) A new test set for validating predictions of protein-ligand interaction, *Proteins: Structure, Function, and Bioinformatics* 49, 457-471.
171. Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wahlby, A., and Jones, T. A. (2004) The Uppsala Electron-Density Server, *Acta Crystallographica Section D-Biological Crystallography* 60, 2240-2249.
172. Dermoun, Z., Foulon, A., Miller, M. D., Harrington, D. J., Deacon, A. M., Sebban-Kreuzer, C., Roche, P., Lafitte, D., Bornet, O., Wilson, I. A., and Dolla, A. (2010) TM0486 from the Hyperthermophilic Anaerobe *Thermotoga maritima* is a Thiamin-binding Protein Involved in Response of the Cell to Oxidative Conditions, *Journal of Molecular Biology* 400, 463-476.
173. Tran, A. T., Cergol, K. M., Britton, W. J., Bokhari, S. A. I., Ibrahim, M., Laphorn, A. J., and Payne, R. J. (2010) Rapid assembly of potent type II dehydroquinase inhibitors via "Click" chemistry, *Medchemcomm* 1, 271-275.
174. Tran, A. T., Cergol, K. M., West, N. P., Randall, E. J., Britton, W. J., Bokhari, S. A. I., Ibrahim, M., Laphorn, A. J., and Payne, R. J. (2011) Synthesis and evaluation of potent ene-yne inhibitors of type II dehydroquinases as tuberculosis drug leads, *Chemmedchem* 6, 262-265.
175. Mitsuhashi, S., and Davis, B. D. (1954) Aromatic Biosynthesis .12. Conversion of 5-Dehydroquinic Acid to 5-Dehydroshikimic Acid by 5-Dehydroquinase, *Biochimica Et Biophysica Acta* 15, 54-61.
176. Chaleff, R. S. (1974) The inducible quinate-shikimate catabolic pathway in *Neurospora crassa*: induction and regulation of enzyme synthesis, *J Gen Microbiol* 81, 357-372.

177. Giles, N. H., Partridge, C. W., Ahmed, S. I., and Case, M. E. (1967) The occurrence of two dehydroquinases in *Neurospora crassa*, one constitutive and one inducible, *Proc Natl Acad Sci U S A* 58, 1930-1937.
178. Kleanthous, C., Deka, R., Davis, K., Kelly, S. M., Cooper, A., Harding, S. E., Price, N. C., Hawkins, A. R., and Coggins, J. R. (1992) A comparison of the enzymological and biophysical properties of two distinct classes of dehydroquinase enzymes, *The Biochemical journal* 282 ( Pt 3), 687-695.
179. Gourley, D. G., Shrive, A. K., Polikarpov, I., Krell, T., Coggins, J. R., Hawkins, A. R., Isaacs, N. W., and Sawyer, L. (1999) The two types of 3-dehydroquinase have distinct structures but catalyze the same overall reaction, *Nature Structural Biology* 6, 521-525.
180. Roszak, A. W., Robinson, D. A., Krell, T., Hunter, I. S., Fredrickson, M., Abell, C., Coggins, J. R., and Laphorn, A. J. (2002) The structure and mechanism of the type II dehydroquinase from *Streptomyces coelicolor*, *Structure* 10, 493-503.
181. White, P. J., Young, J., Hunter, I. S., Nimmo, H. G., and Coggins, J. R. (1990) The purification and characterization of 3-dehydroquinase from *Streptomyces coelicolor*, *The Biochemical journal* 265, 735-738.
182. Bottomley, J. R., Clayton, C. L., Chalk, P. A., and Kleanthous, C. (1996) Cloning, sequencing, expression, purification and preliminary characterization of a type II dehydroquinase from *Helicobacter pylori*, *Biochemical Journal* 319, 559-565.
183. Payne, R. J., Peyrot, F., Kerbarh, O., Abell, A. D., and Abell, C. (2007) Rational design, synthesis, and evaluation of nanomolar type II dehydroquinase inhibitors, *Chemmedchem* 2, 1015-1029.
184. Gonzalez-Bello, C., and Castedo, L. (2007) Progress in type II dehydroquinase inhibitors: From concept to practice, *Medicinal Research Reviews* 27, 177-208.
185. Bailey, S. (1994) The Ccp4 Suite - Programs for Protein Crystallography, *Acta Crystallographica Section D-Biological Crystallography* 50, 760-763.
186. Singh, S., Stavriniades, J., Christendat, D., and Guttman, D. S. (2008) A phylogenomic analysis of the shikimate dehydrogenases reveals broadscale functional diversification and identifies one functionally distinct subclass, *Molecular Biology and Evolution* 25, 2221-2232.
187. Singh, S., Korolev, S., Koroleva, O., Zarembinski, T., Collart, F., Joachimiak, A., and Christendat, D. (2005) Crystal structure of a novel shikimate dehydrogenase from *Haemophilus influenzae*, *Journal of Biological Chemistry* 280, 17101-17108.

## Appendices (Appendix 1)

Legend key

e.d.a = electron density absent

- = protein structure not subjected to pharmacophore searching

Rows highlighted green = protein structures subjected to pharmacophore searching

Rows highlighted yellow = protein structures without electron density

Rows highlighted blue = protein structures visualized in Coot for viewing unknown ligands along with electron density

PDB code	visual description of unknown ligand(s) and binding domain of the target protein	structure selected for pharmacophore searching	comments about pharmacophore search hits
1FSC	no electron density available for the protein structure	e.d.a	
1O5U	cluster of electron density , may be coordinating water molecules	-	
1O5Z	Small ligand with low electron density in the binding site of protein	-	
1OR8	no electron density available for the protein structure	e.d.a	
1ORI	no electron density available for the protein structure	e.d.a	
1S01	no electron density available for the protein structure	e.d.a	
1SVV	metal site with disordered electron density	-	
1TVF	medium size ligand with good electron density in the binding domain	⇒	among other hits citric acid and its derivatives efficiently satisfy the model
1UEH	ligand with good electron density only in B chain, but not in the active site	-	
1VJ0	no electron density available for the protein structure	e.d.a	
1VJ2	ligand with partially disordered electron density, not covalently bound to the manganese atom in the active site, may be water molecules	-	
1VJL	small ligand with disordered electron density and not in the binding site	-	
1VK2	very small ligand, may be a water molecule in active site	-	
1VK8	well defined ligand in the active site with good electron density	⇒	thiamine and its derivatives satisfy the pharmacophore
1VK9	ligand with poor resolution, probably metal ion with coordinating waters molecules	-	
1VKE	no electron density available for the protein structure	e.d.a	
1VKM	medium size ligand with well defined electron density close to the metal centre in the active site	⇒	As a unique hit D-erythritol 4-phosphate logically fits both the electron density and the pharmacophore model
1VKY	ligand not with well defined electron density	-	
1VL0	ligand with good electron density but present out side the active site of the protein and only in chain A, and absent in chain B	⇒	the stereo chemistry fits reasonably to aldehydo-D-allose as a hit which is not in the crystallization conditions, possibly a potential candidate as a true ligand
1VMG	no electron density available for the protein structure	e.d.a	



1VP4	very small ligand,probably a water molecule in active site	-	
1VP8	ligand with not a well defined electron density, may be a metal ion coordinating with water molecules	-	
1VPY	small ligand covalently bounded to cysteine out side the active site of the protein	-	
1VQ0	no electron density available for the protein structure	e.d.a	
1VQZ	ligand with good electron density ,covalently bonded to glutamate, but not in the active site of the protein	-	
1VR3	no electron density available for the protein structure	e.d.a	
1VRA	no electron density available for the protein structure	e.d.a	
1VRM	small ligand with disordered electron density, with in the active site	-	
1YBA	no electron density available for the protein structure	e.d.a	
1YQ3	17 very small ligands scattered far a part from each other , majority at the periphery of the protein, probably water molecules	-	
1Z8H	well defined 5 member ring ligand covalently bonded to serine(in all A,B,C and D chains ),in the active site	-	
1ZKG	small ligand with low electron density in a more hydrophobic pocket of the protein with provision of very few interactions ,probably a small fatty acid or PEG molecule in binding pocket	-	
2A06	17 small ligands scattered far a part , majority at the periphery, may be these are water molecules	-	
2A0N	large ligand covalently bound to cysteine in the binding pocket, can be peptide with disulphide linkage to cysteine residue of the protein	-	
2A6A	ligand with reasonable electron density in the binding site, appears like sausage hydrophobic group	-	
2AAM	reasonably long ligand with fair electron density in the binding pocket, found in all A,B,C,D,E and F chains of the protein	⇒	beta-D-galacturonic acid seems a logical hit by forming additional H-bond interactions at position 1 and 4
2AJ6	small ligand with fair electron density, looks like a water molecule in the binding site	-	
2F1L	two ligands with poor electron density , only in chain A	-	
2F46	poor density small ligand at the interface of chain B only and not in A	-	
2F4P	medium size ligand with poor electron densities in Chain A, B,C and D	-	
2F6R	small ligand with reasonably good electron density in the active site	⇒	among hits citramalic acid is convincingly satisfying the model
2FBW	17 small ligands scattered far a part , majority at the periphery, may be these are water molecules	-	
2FCL	small size ligand with disordered electron density	-	
2FTR	small ligand in the binding site with reasonable electron density	⇒	Few good hits like homo-serine, L-leucine, L-homoserine, creatine, L-aspartic acid and L-asparagine comply with the pharmacophore model.
2FTZ	two ligands in two binding sites in the same chain with poor electron density	-	
2G8L	small ligand in each chain(A and B) with poor electron density	-	
2GFG	small ligand with poor electron density in each binding pocket(A,B and C chain)	-	
2GNO	small ligand in the binding pocket with reasonable electron density , possibly water	-	

2GVI	molecules small ligand with good electron density, covalently bounded to cysteine ,near to the zinc metal ion in the active site	-	
2GVK	very small ligand, possibly a metal ion.	-	
2H1N	small ligand with poor electron density	-	
2H88	17 small ligands scattered far a part , majority at the periphery, may be these are water molecules	-	
2H89	no electron density available for the protein structure	e.d.a	
2HBW	small ligand doesn't match the electron density.	-	
2HSZ	small sausage shaped ligand with low electron density, may be a PEG molecule	-	
2I5N	no electron density found around the ligand.	-	
2I8D	small ligand , may be water molecules, no bonding electron density visible between the ligand fraction	-	
2IG6	ligand with good electron density in the binding pocket, looks like nitro benzene only in chain A and not in chain B	-	
2ITB	in Chain B ligand with reasonable electron density, fraction of ligand covalently bounded to Fe, but absent in Chain A.	-	
2NQL	poor electron density ligands looks like nitro benzene, located in the binding pocket, both in chain A and B	-	
2O08	long ligand with fair electron density in the binding site of chain A, but absent in chain B	-	
2O7T	two separate long ligands some what like fatty acids, with fair density in the binding pocket. Might be a PEG molecule	-	
2OC5	long sausage like ligand with good electron density, covalently linked to Fe.	-	
2OH1	small parts of ligands in Chain A,B and D and absent in C, may be 3-4 water molecules	-	
2OLM	no electron density available	e.d.a	
2OPK	ligand like 5 member ring, with reasonable electron density ,located well inside the binding pocket in all A,B,C and D chains, possibly a imidazole ring.	-	
2OUW	medium size ligand with fair electron density in the binding pocket in both A and B chains	-	
2OWN	ligand like medium size fatty acid in the binding site with good electron density in both chains	⇒	L-arginine, N-(2-aminoethyl)butane-1,4-diamine and hexyl isocyanate were the best hits, satisfying both the electron density and the pharmacophore model
2OZH	small ligand with good electron density in the binding pocket, possibly a water molecule	-	
2PBL	good electron density ligand like benzamidine in all A,B,C and D chains of the protein	-	
2PCS	very big ligand with low electron density, located in the binding site	-	
2PEB	small ligand with good electron density in the active site near Zn	-	
2PNK	ligand with good electron density, looks like a phosphate in the binding pocket ,present only in chain B and absent among chains A-L.	-	
2PY6	medium size ligand with good electron density in the active site of the protein	⇒	best hits included some amino acids and sugars
2Q02	very small ligand with low electron density covalently bound to Zn in the active site	-	
2Q0Y	small ligand with low electron density in the binding site	-	
2Q83	medium size ligand with low electron density, present in the binding site in both chains	-	
2Q9K	fairly large ligand with disordered electron density in the binding pocket	-	
2QDR	cluster of electron density probably reminiscent of metal ion and coordinating water	-	

2QE8	molecules, no bonding electron density present between the ligand atoms	-	
2QGG	low electron density small ligand, looks like a fatty acid or PEG molecule in binding pocket	-	
	long ligand at the periphery of the protein, bonding electron density absent between the atoms	-	
2QIW	medium size ligand with low electron density, present in chain A but absent in chain B	-	
2RAU	very large ligand with good atomic and bonding electron density ,well inside the binding pocket ,some what like long chain fatty acid	-	
2W87	well defined ligand , with good atomic and bonding electron density, covalently attached to glucuronic acid in both chains at the periphery of the protein	-	
2WQY	12 small ligands with poor electron density scattered far a part from each other , majority at the periphery, probably water molecules	-	
2WSC	protein structure comprised of 19 chains with poor electron density	-	
2WSE	protein structure comprised of 19 chains with poor electron density	-	
2WSF	protein structure comprised of 19 chains with poor electron density	-	
2WTL	no electron density available for the protein structure	e.d.a	
3ARC	structure comprised of 52 chains, with low electron density	-	
3B9Y	small ligand, electron density not matching with the ligand atoms	-	
3B9Z	small ligand with low electron density in the binding site	-	
3BI7	no electron density available for the protein structure	e.d.a	
3BN8	medium size sausage shape ligand with good atomic and bonding electron density in the binding site in chain A but absent in chain B	-	
3BOG	no electron density available for the protein structure	e.d.a	
3C5E	medium size ligand with low electron density in the binding pocket	-	
3C5I	Very small ligand with good electron density blob, probably a metal ion. present in chain B,C and D but absent in chain A	-	
3CNX	medium size ligand with low electron density in the binding pocket in all chains, but bonding electron density absent between the atoms	-	
3CSW	medium size ligand with low electron density, present in the binding pocket of all the chains but bonding electron density absent between the atoms	-	
3CT8	very long ligand with good electron density and covalently attached to Zn in the active site	⇒	pharmacophore not pulling out any ligands
3CWB	protein structure comprised of 20 chains, with poor electron density	-	-
3D1C	very small scattered ligand with poor electron density in the binding site, probably scattered water molecules	-	
3D82	ligand like benzamidine with low electron density, covalently linked to Ni in the active site in all A,B,C,D and E chains of the protein	-	
3D9R	no electron density available for the protein structure	e.d.a	
3DDL	10 fragments of medium and large size ligands with low electron density in both chains	-	
3E5D	a very large ligand with low electron density in the binding pocket	-	
3E8O	small ligand with low electron density in the binding pocket in both chains but bonding electron density absent between the ligand atoms	-	
3E8V	very small ligand with good electron density in the binding site, may be an ion or metal ion	-	
3EBT	fragmented ligand with low electron density in the binding pocket, may be water molecules	-	
3ECF	ligand like benzamidine with low electron density, present in the active site in all A,B,C	-	

3ED4	and D chains of the protein small ligand with low electron density present in the binding pocket and in the periphery in all A, B, C and D chains of the protein	-	
3EJK	very small ligand with good electron density in the binding pocket, may be a water molecule	-	
3EJV	ligand like benzamidine with low electron density in the binding pocket of the protein	-	
3EK3	no electron density available for the protein structure	e.d.a	
3EZ0	very long ligand with low electron density in the binding pocket in all A,B,C and D chains, probably a PEG molecule	-	
3EZU	medium size ligand with good electron density in the binding pocket of the protein	⇒	2-ethylhexan-1-ol and diethylaminoethanol fit the electron density and the model and can be potential ligands of the protein
3F0H	no electron density available for the protein structure	e.d.a	
3F7S	no electron density available for the protein structure	e.d.a	
3F7W	no electron density available	-	
3F7X	very small ligand with good electron density in the binding pocket at two different sites, probably an ion	-	
3FDE	small scattered ligands in and around protein with good electron density, in all chains, probably water molecules	-	
3FF0	ligand like benzamidine with poor electron density, present at two binding sites of chain A but absent in chain B	-	
3FGV	fragmented ligand with good electron density in the binding site in both chains	-	
3FGY	very small ligand with poor electron density in the binding site in both chains, can be an ion	-	
3FH1	medium size ligand with good electron density in the binding pocket	⇒	hits like pantothenate, D-galactonic acid and L-citrulline reasonably fitted the pharmacophore model
3FKA	medium size ligand with good electron density in the binding site in all A,B,C and D chains of the protein	-	
3FLJ	small ligand with good electron density in the binding pocket	-	
3FSD	fragmented ligand with poor electron density in the binding pocket	-	
3G16	medium size ligand with good electron density in the binding pocket in both chains	-	
3G23	medium size ligand with low electron density in the binding site of chain A, but absent in chain B	-	
3GBH	medium size ligand with low electron density ligand in the binding site, in all A,B,C and D chains of the protein	-	
3GBN	no electron density for protein.	-	
3GE5	benzamidine like ligand with good electron density in the binding pocket in both chains	-	
3GGD	medium size ligand with good electron density in the binding site	-	
3GI7	medium size ligand with low electron density in the binding site in both chains	-	
3GIW	ligand like benzene ring with low electron density in the binding pocket of the protein	-	
3GQQ	fragmented ligand with good electron density in the active site present in all A-F chains of the protein	-	
3GR3	medium size fragmented ligand with good electron density in the binding site in both chains	-	
3GWR	benzamidine like ligand in the binding site with good electron density in both chains	-	
3GZA	fragmented medium size ligand with low electron density in chain B only but absent in chain	-	

	A	
3GZI	medium size fragmented ligand with low electron density in the binding site	-
3GZR	benzamidine like fragmented ligand with good electron density in the binding site in both chains	-
3H1H	protein comprised of 20 chains with poor electron density	-
3H1I	protein comprised of 20 chains with poor electron density	-
3H1J	protein comprised of 20 chains with poor electron density	-
3H1K	protein comprised of 20 chains with poor electron density	-
3H1L	protein comprised of 20 chains with poor electron density	-
3H3H	medium size ligand with low electron density in the binding pocket in both chains	-
3H4Q	fragmented medium size ligand with low electron density in the binding site	-
3H5I	small ligand with low electron density in the binding site in both chains	-
3H7A	medium size ligand with low electron density in the binding pocket in chains C and D only, but absent in chains A and B, probably a PEG molecule	-
3HFT	small ligand with low electron density in the active site near the metal Zn.	-
3HL1	fragmented ligand with low electron density in the binding site in both chains	-
3HM4	medium size ligand with low electron density in the binding pocket in chain A only and absent in chain B	-
3HMZ	medium size fragmented ligand with poor electron density in the binding site	-
3HOI	small ligand with poor electron density in the periphery of the protein.	-
3HRG	fragmented ligand with low electron density in the binding site, probably water molecules	-
3HX8	medium size ligand with good electron density in the active site in all A,B,C and D chains of the protein	-
3IOY	medium size ligand with good electron density in the active site in all A,B,C and D chains of the protein	-
3IEH	medium size ligand with fair electron density in the active site near the Zn, probably a glycerol molecule	-
3JR1	medium size ligand with low electron density in the binding pocket of both chains of the protein	-
3JTW	medium size ligand with low electron density in the binding site in chain A only but absent in chain B, probably a PEG molecule	-
3JVG	fairly long two ligands with poor electron density at two different sites in the binding site in chains A and B only but absent in chains C and D	-
3KL7	small ligand with poor electron density covalently linked to Zn in the active site	-
3KS6	small ligand with good electron density, covalently linked to Mg in the active site in all A, B, C and D chains, probably an acetate molecule	-
3KTC	good density, medium size fragmented ligand, covalently linked to Fe in the active site in both chains	-
3KTS	medium size ligand with poor electron density in the binding site in all chains A-H of the protein	-
3KWK	very small ligand, with fair electron density, probably a single atom ion or water molecule	-
3KY8	medium size fragmented ligand with low electron density, in the binding site in both chains	-
3L12	medium size ligand with good electron density, covalently linked to Mg in the active site in	-

3L2N	both chains small ligand with low electron density covalently linked to Ca ion in the active site in both chains of the protein	-	
3L49	medium size ligand with low electron density in the binding pocket in all A,B,C and D chains	-	
3L60	ligand like benzamidine with good electron density in the active site	-	
3L74	protein structure comprised of 26 chains with poor electron density	-	
3L75	protein structure comprised of 26 chains with poor electron density	-	
3LB5	benzamidine like ligand with good electron density in the binding pocket in all A,B,C and D chains of the protein	-	
3LOT	small ligand with low electron density covalently linked to Zn in the active site in all chains A,B,C and D ,probably a water molecule	-	
3LOU	small fragmented ligand with poor electron density in the binding site in A chain only and absent in B,C and D chains	-	
3LWU	medium size ligand with low electron density, covalently linked to Zn in the active site	-	
3LYG	ligand like benzamidine with low electron density in the active site	-	
3M5K	fragmented ligand with poor electron density in the binding site in both chains	-	
3MHO	medium size ligand with low electron density in the binding site , probably a PEG molecule	-	
3MPR	fragmented ligand with low electron density in the binding site of the protein in chain A only and absent in chain B,C and D	-	
3MST	medium size ligand with low electron density in the binding site, probably a PEG molecule	-	
3N5L	medium size ligand with low electron density in the binding site in both chains	-	
3N7O	medium size ligand with good electron density in the binding site.	-	
3NDO	medium size ligand with good electron density in both chains	-	
3NE7	small fragmented ligand with low electron density in the binding site	-	
3NF6	no electron density available for the protein structure	e.d.a	
3NF7	no electron density available for the protein structure	e.d.a	
3NF8	no electron density available for the protein structure	e.d.a	
3NF9	no electron density available for the protein structure	e.d.a	
3NG3	medium size ligand with good electron density in the binding site in all Chain A-D	-	
3NNR	comparatively long ligand with good electron density in the binding site of the protein	⇒	Among other long chain fatty acids as hits, palmitic acid fits the model in the best way
3NO4	very small ligand with low electron density near the Ni atom in the active site in all A-C chains , probably an ion	-	
3NRB	benzamidine like good electron density ligand in the binding site in chain A only and absent in chains B,C and D	-	
3NRJ	no electron density available for the protein structure	e.d.a	
3OE4	no electron density available for the protein structure	e.d.a	
3OE5	no electron density available	e.d.a	
3OF4	fragmented ligand with good electron density, located in the periphery of the protein in Chain A only and absent in chains B,C and D	-	
3OKH	no electron density available for the protein structure	e.d.a	
3OKI	no electron density available for the protein structure	e.d.a	
3OLF	no electron density available for the protein structure	e.d.a	

3OLQ	no electron density available for the protein structure	e.d.a	
3OMK	no electron density available for the protein structure	e.d.a	
3OOF	no electron density available for the protein structure	e.d.a	
3OOK	no electron density available for the protein structure	e.d.a	
3OP7	purine ring like ligand with good electron density in the binding site of the protein	-	
3OS5	electron density not matching with the atoms positions, ligand in chain A only and not in chain B	-	
3OYV	very small ligand, with good electron density in the binding pocket, probably a water molecule	-	
3OZR	no electron density available for the protein structure	e.d.a	
3OZS	no electron density available for the protein structure	e.d.a	
3OZT	no electron density available for the protein structure	e.d.a	
3P6K	Purine ring like ligand with good electron density in the binding site in chain B but absent in chain A	⇒	among other hits, pyridoxilactone fits the model comparatively in a better manner
3PCV	very small ligands with poor electron density , scattered far a part from each other, in the periphery of the protein	-	
3PIK	medium size ligand with poor electron density in the binding site , covalently linked to cysteine of the protein	-	
3PM9	fragmented ligand with poor electron density in the binding site in all chains A-F of the protein	-	
3PMI	medium size ligand with low electron density, present in the binding site in C and D chains only, but absent in chains A and B	-	
3PPL	fragmented ligand with low electron density in the binding site in both chains of the protein	-	
3QC0	medium size ligand with fair electron density in the active site covalently linked to Zn ion	-	
3QK8	benzamidine like ligand with fair electron density in the binding site in all chains A-F, except B of the protein	-	
3QQV	no electron density available	-	
3QUA	Very small ligand with good electron density in the binding pocket, possibly a water molecule or a metal ion	-	
3QUQ	medium size ligand with good electron density in the active site near the Mg ion	-	
3QYP	good electron density small ligand in the binding site in chain B only and not in A	-	
3RH2	very long ligand with good electron density in the binding site	-	

## Legend key

e.d.a = electron density absent

- = protein structure not subjected to pharmacophore searching

Rows highlighted green = protein structures subjected to pharmacophore searching

Rows highlighted yellow = protein structures without electron density

Rows highlighted blue = protein structures visualized in Coot for viewing unknown ligands along with electron density